2014-02-07

Digital Forensics Overview

Garfinkel, Simson L.

Monterey, California. Naval Postgraduate School

Digital Forensics

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

Feb 7, 2014

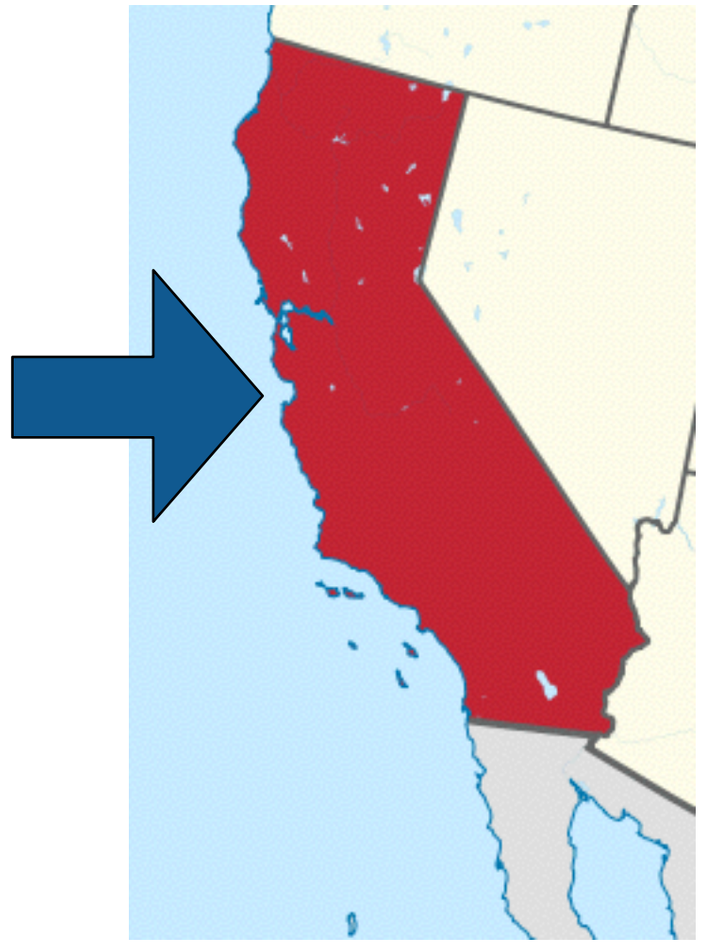
http://www.forensicswiki.org/

http://simson.net/

NPS is the Naval Postgraduate School

Monterey, CA — 1500 students

- US Military & Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)



Graduate Schools of Operational & Information Sciences (GSOIS)

- Computer Science
- Defense Analysis
- Information Sciences
- Operations Research
- Cyber Academic Group

National Capital Region (NCR) Office

- 900 N Glebe (Ballston)/Virginia Tech building
- Courses at Ft. Meade

Digital Evaluation and Exploitation (DEEP): Research in “trusted” systems and exploitation.

“Evaluation”

- Trusted hardware and software
- Cloud computing

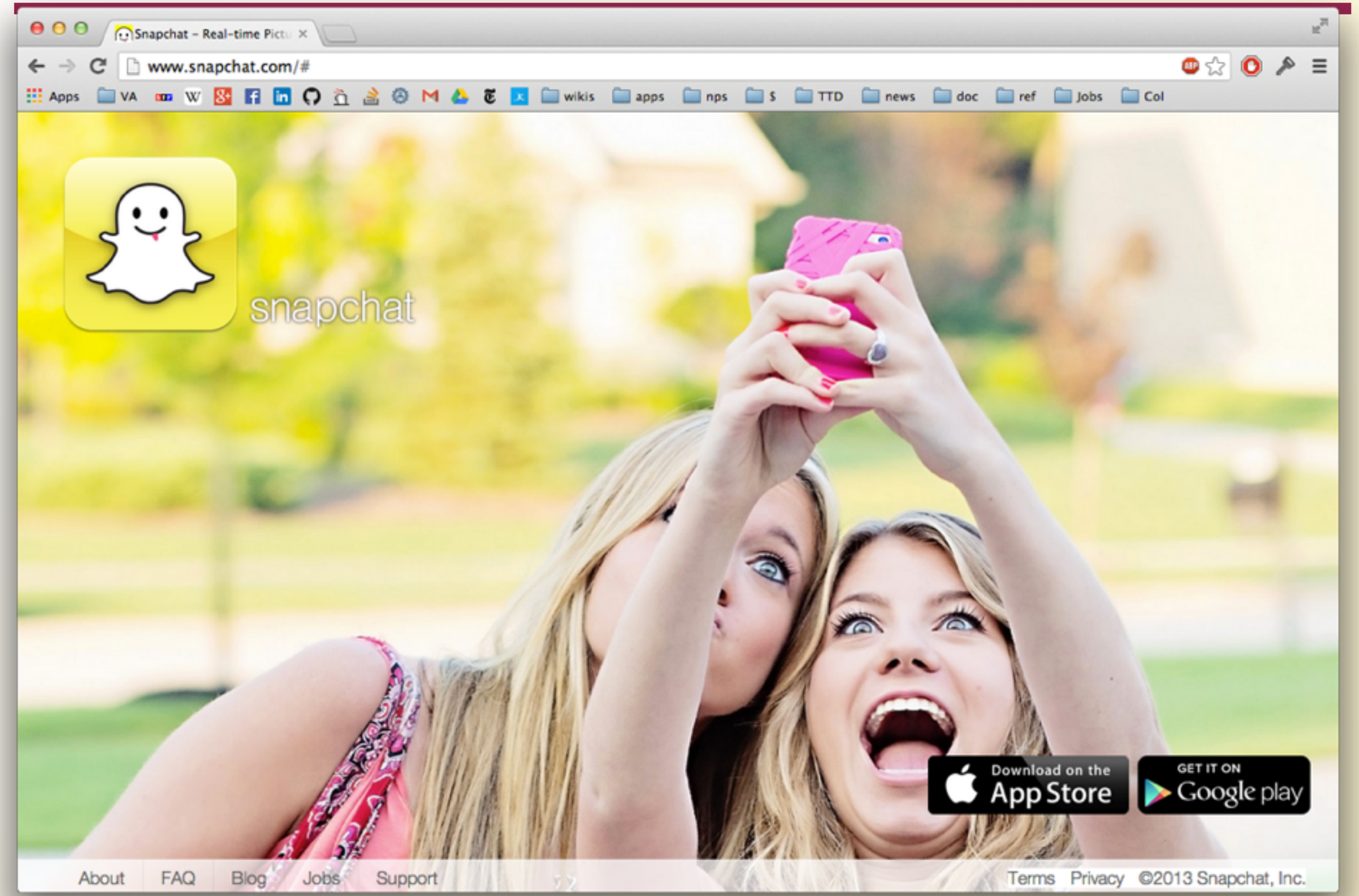
“Exploitation”

- MEDEX — “Media” — Hard drives, camera cards, GPS devices.
- CELEX — Cell phone
- DOCEX — Documents
- DOMEX — Document & Media Exploitation

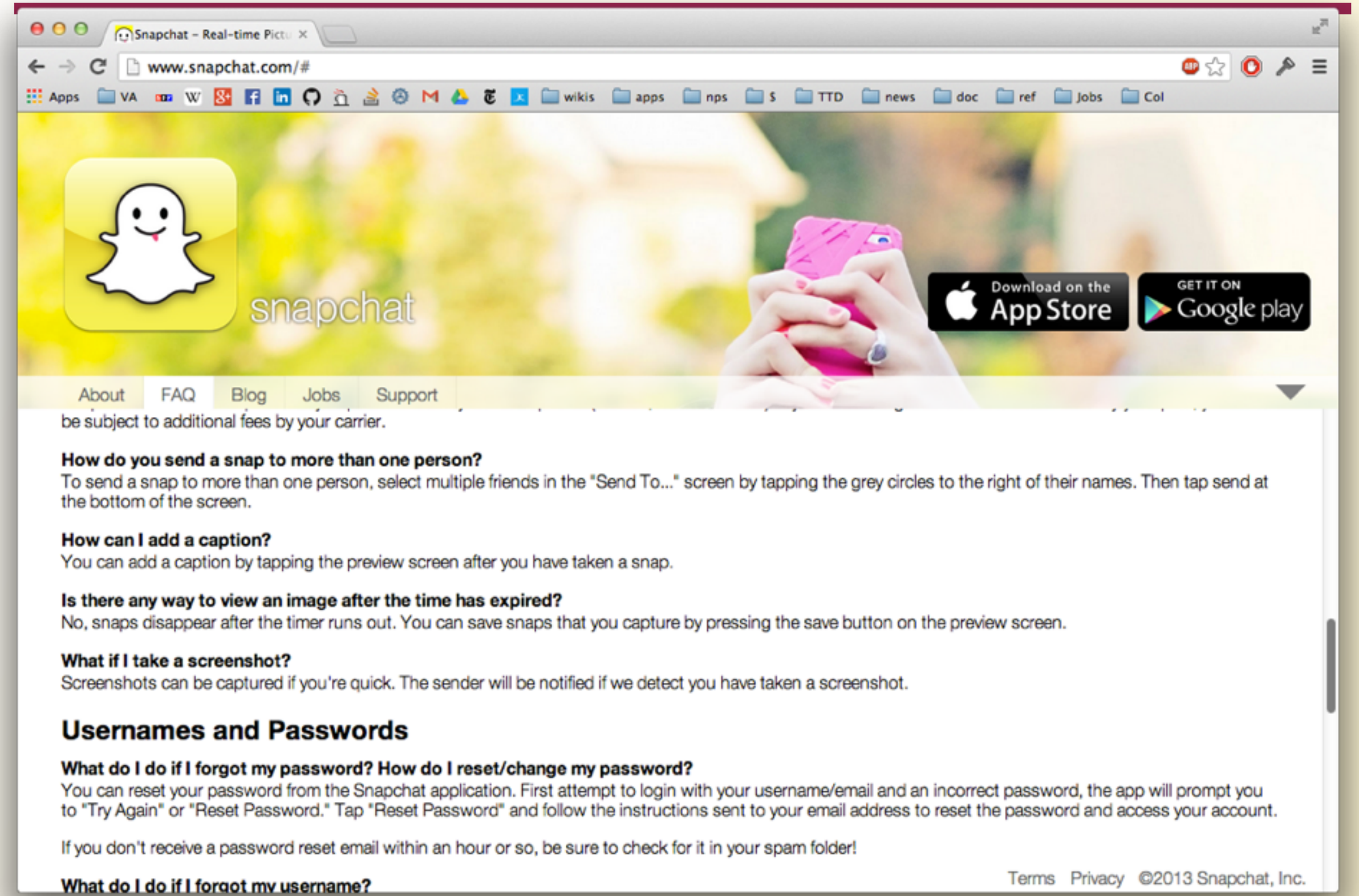
Current Partners:

- Law Enforcement (FBI & Local)
- DHS (HSARPA; Video Games & Insider Threat
- NSF (Courseware development)
- DOD

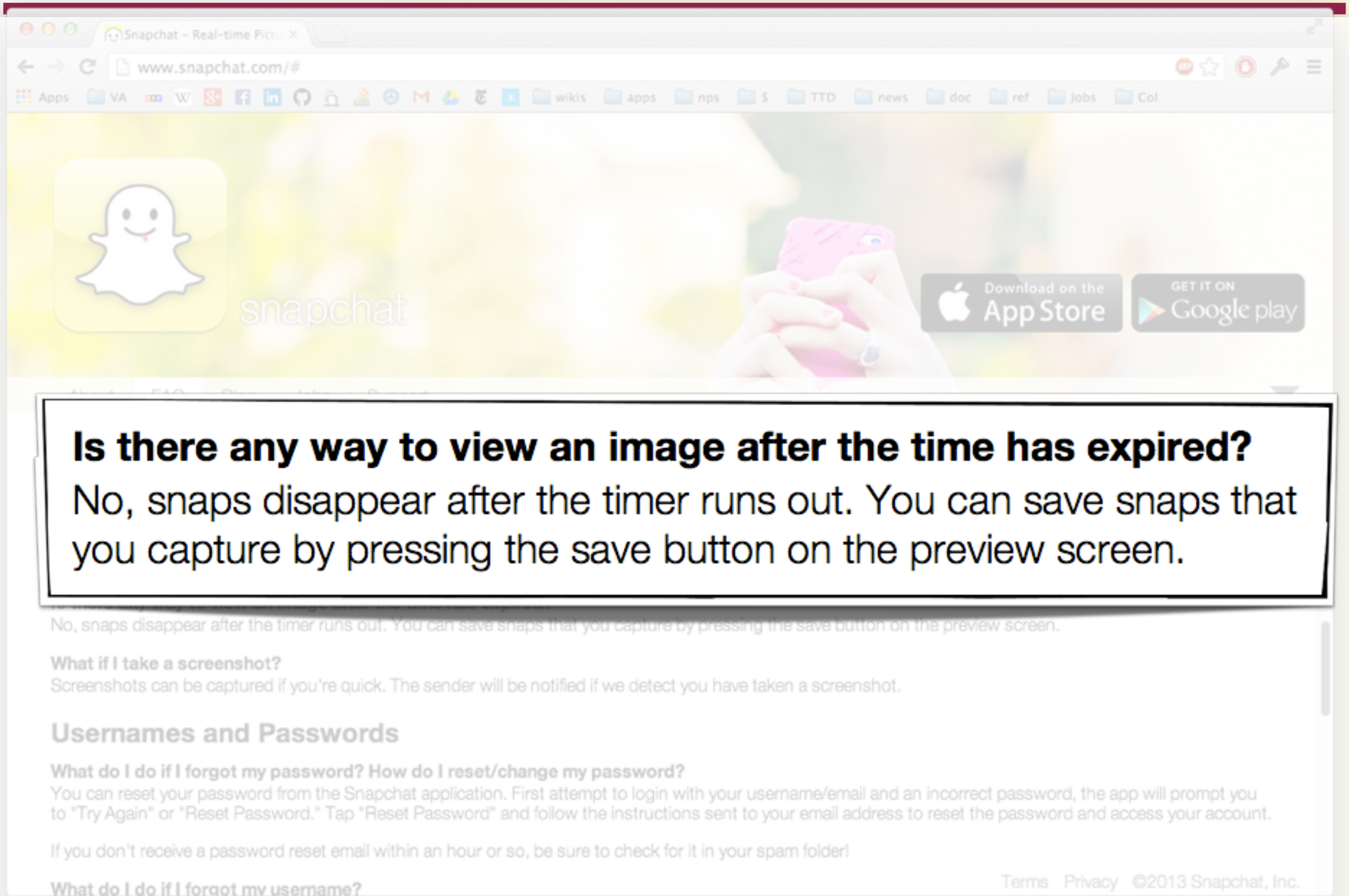
Snapchat is a popular app!



Snapchat promised users that expired images could not be viewed unless “saved.”



Snapchat promised users that expired images could not be viewed unless “saved.”



Is there any way to view an image after the time has expired?

No, snaps disappear after the timer runs out. You can save snaps that you capture by pressing the save button on the preview screen.

OMG! — Expired images not actually deleted. They were just hidden from view.

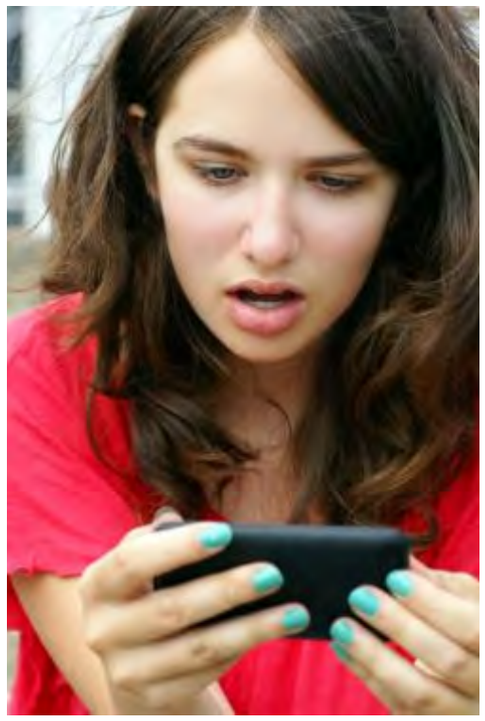
slate | Follow @slate 588K followers | LO

NEWS & POLITICS | TECH | BUSINESS | ARTS | LIFE | HEALTH & SCIENCE | SPORTS | DOUBLE X | PODCASTS

OMG, "Deleted" Snapchat Sexts Can Actually Be Recovered

By Will Oremus | Posted Thursday, May 9, 2013, at 3:56 PM

Share 59 | Like 133 | Tweet 84 | EMAIL | PRINT | COMMENT 46



Snapchat's users shouldn't be shocked to find that their images can be recovered even after they "self-destruct"—but they will be anyway.
Sylvie Bouchard/Shutterstock.com

The premise of Snapchat is simple: Send a photo or short video to a friend, and it will self-destruct after 10 seconds. That way, it won't wind up on the Internet and ruin anyone's reputation, friendships, or career.

Needless to say, that has made it a wildly popular choice for sexting. But Snapchat's appeal goes far beyond that. In an age in which "privacy" and "technology" have become almost antonymous, it has been billed as the anti-Facebook—a communications tool that deletes your data rather than preserving, analyzing, and trading on it. In short, it's supposed to make messaging fun again.

But the app's security has never been ironclad. As the media have repeatedly warned parents, and parents in turn warned their kids, message recipients can still save a compromising image by taking a quick screenshot. But Snapchat tries to mitigate the risk somewhat by automatically notifying the sender when that happens. If someone screenshots you, it's a virtual slap in the face. If they don't, you can assume you're in the clear.

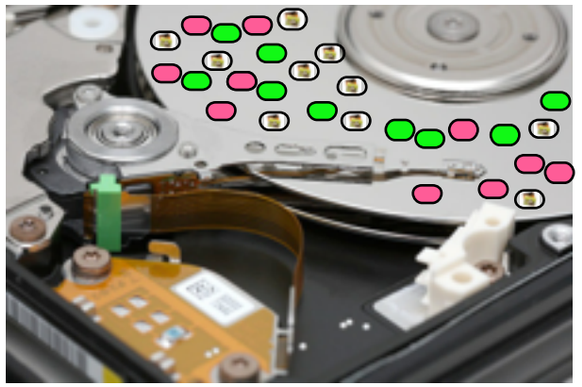
Except that apparently you can't. KSL-TV in Utah reports that an Orem-based firm called Decipher Forensics has figured out a way to recover supposedly deleted images from the recipient's phone. The process isn't simple: 24-year-old Decipher forensics examiner Richard Hickman told the network that it takes him about six hours, on average, to image the phone's data. So far he can only do it with Android devices, though he's working on doing the same for iOS. But his firm is now offering to perform the recovery procedure for anyone who wants it, from parents [illegible] $500

This talk presents today's digital forensic challenges and presents approaches for solving the triage problem.

Introducing digital forensics

Today's digital forensics challenges

Approaches for solving the triage problem



Digital information is pervasive in today's society.

Many potential sources of digital information:

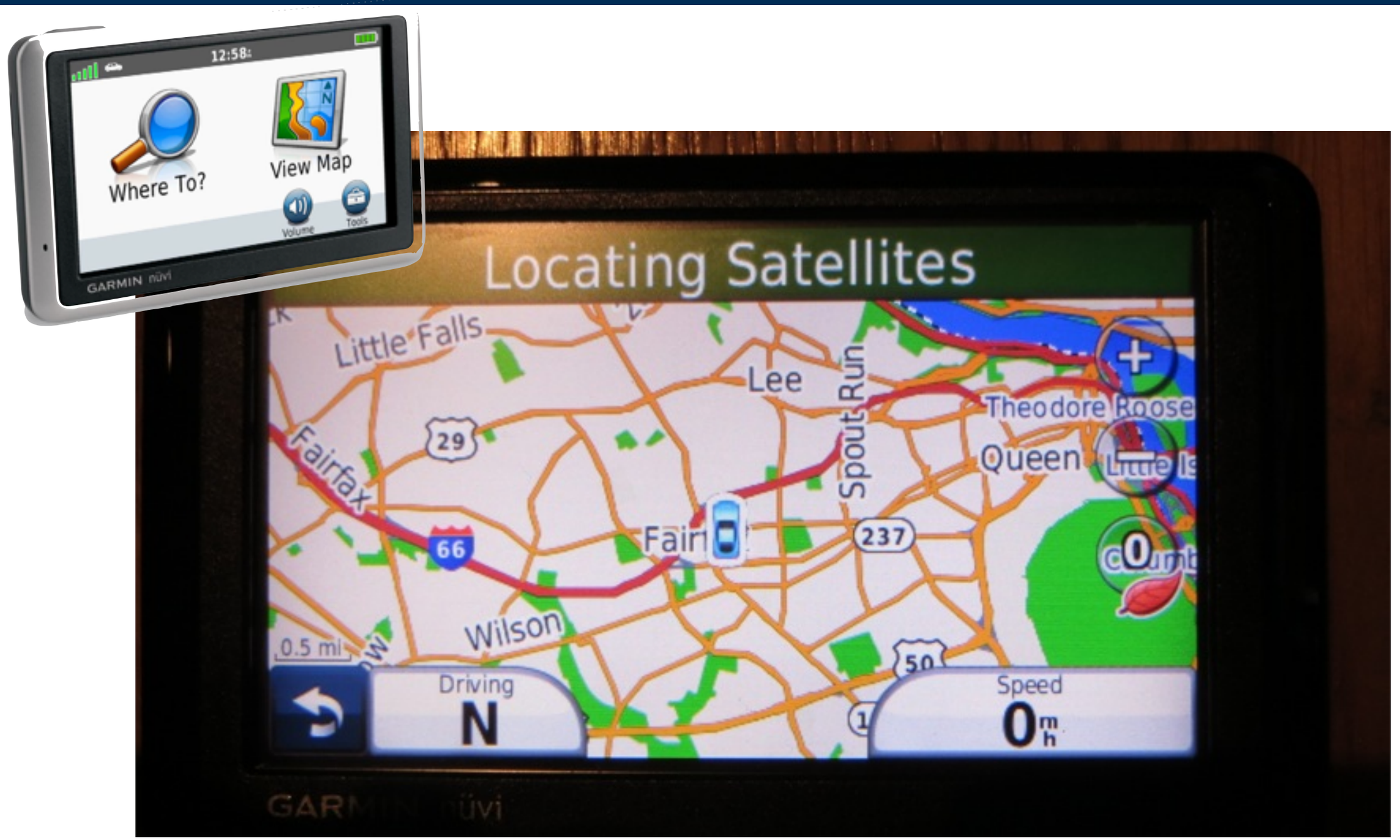
- Desktops; Laptops
- Tablets; Cell Phones
- Internet-Based Services
- Cars

Users of forensic tools have many different goals:

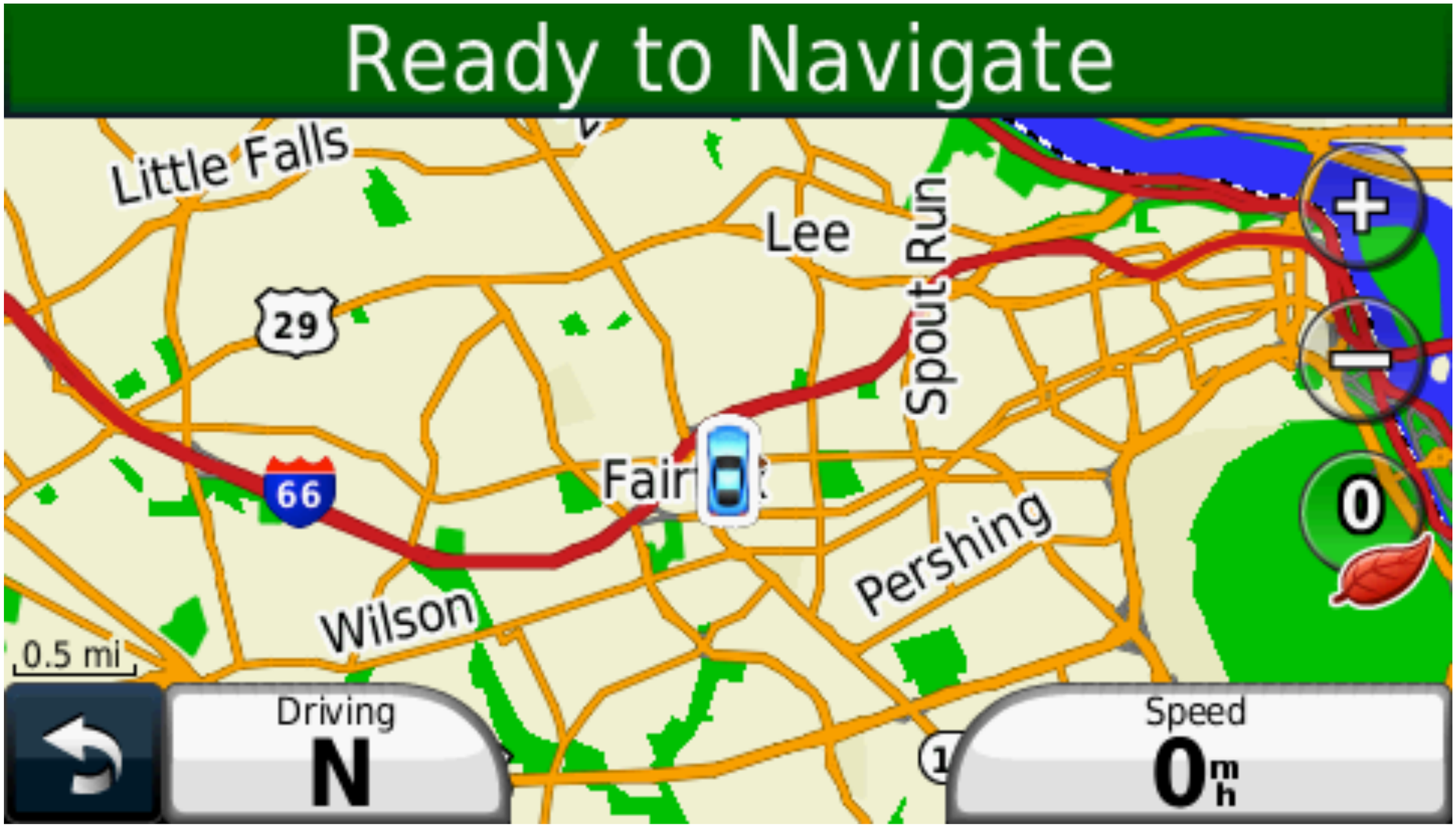
- Document a conspiracy (stock fraud; murder-for-hire)
- Investigation, intelligence & analysis
- Establish possession of contraband information (images; movies)
- Recover "lost" information
- Understand and correct privacy leaks

Many devices record information in non-obvious ways.

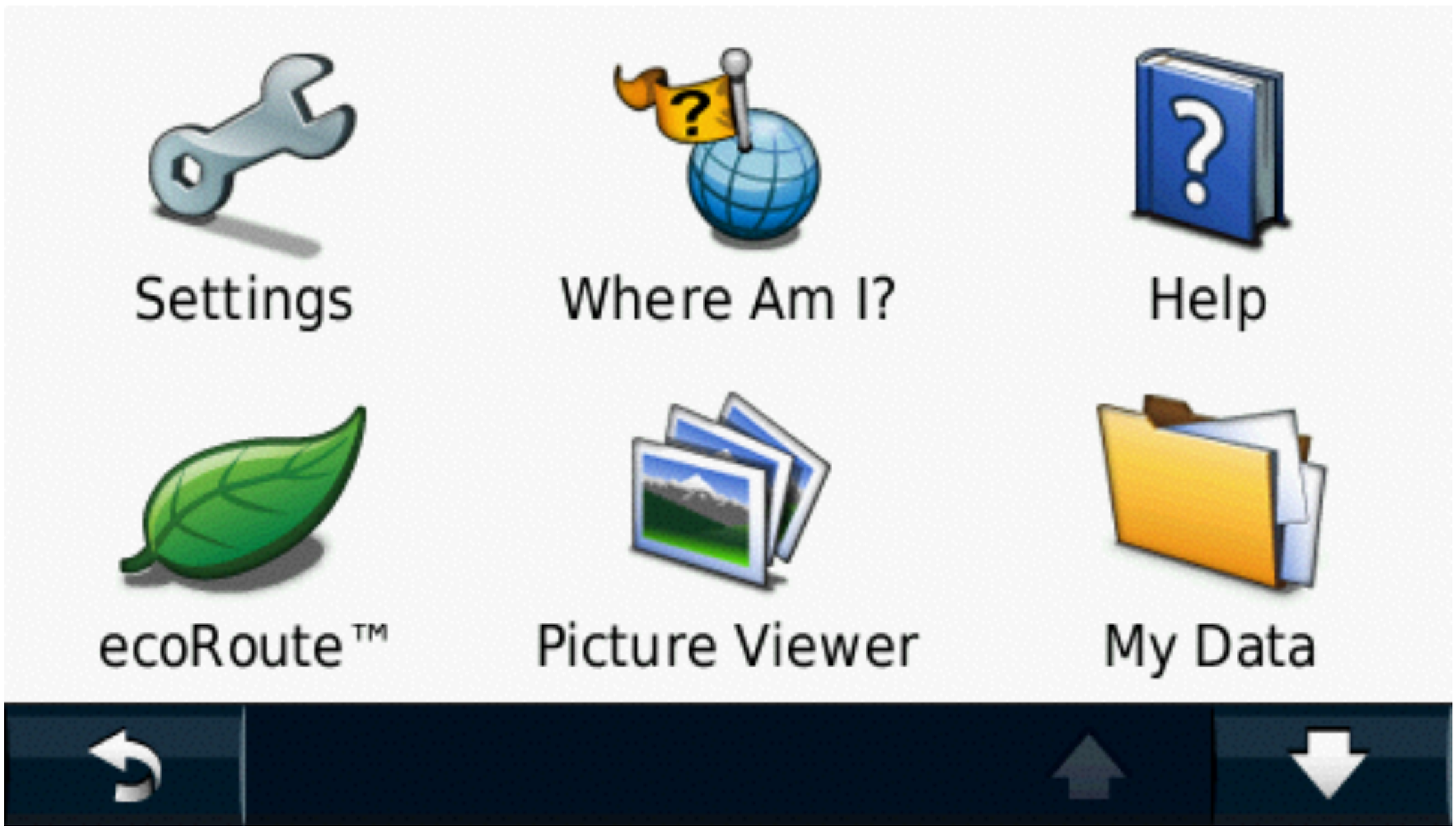
Many devices record information in non-obvious ways.

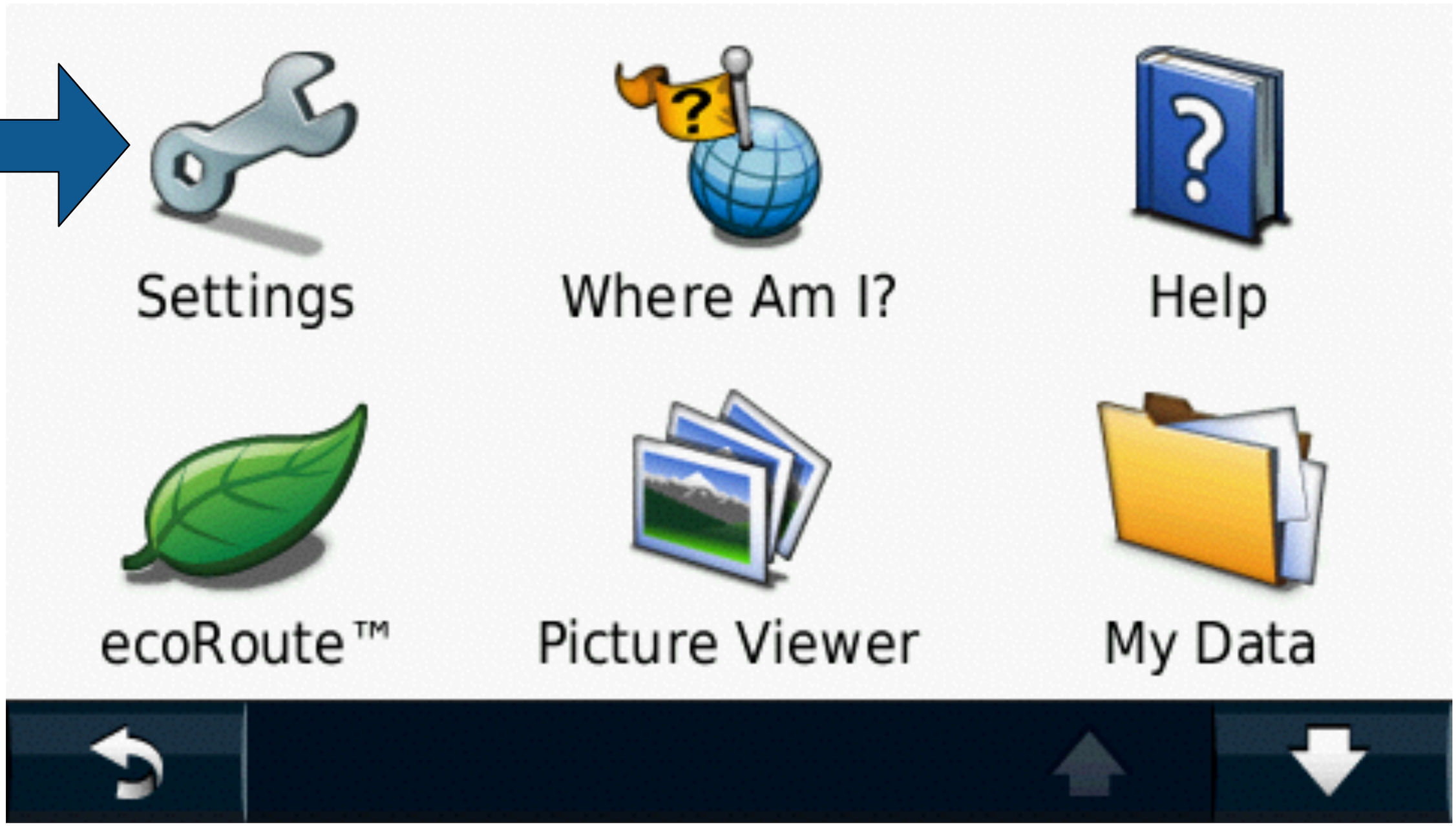


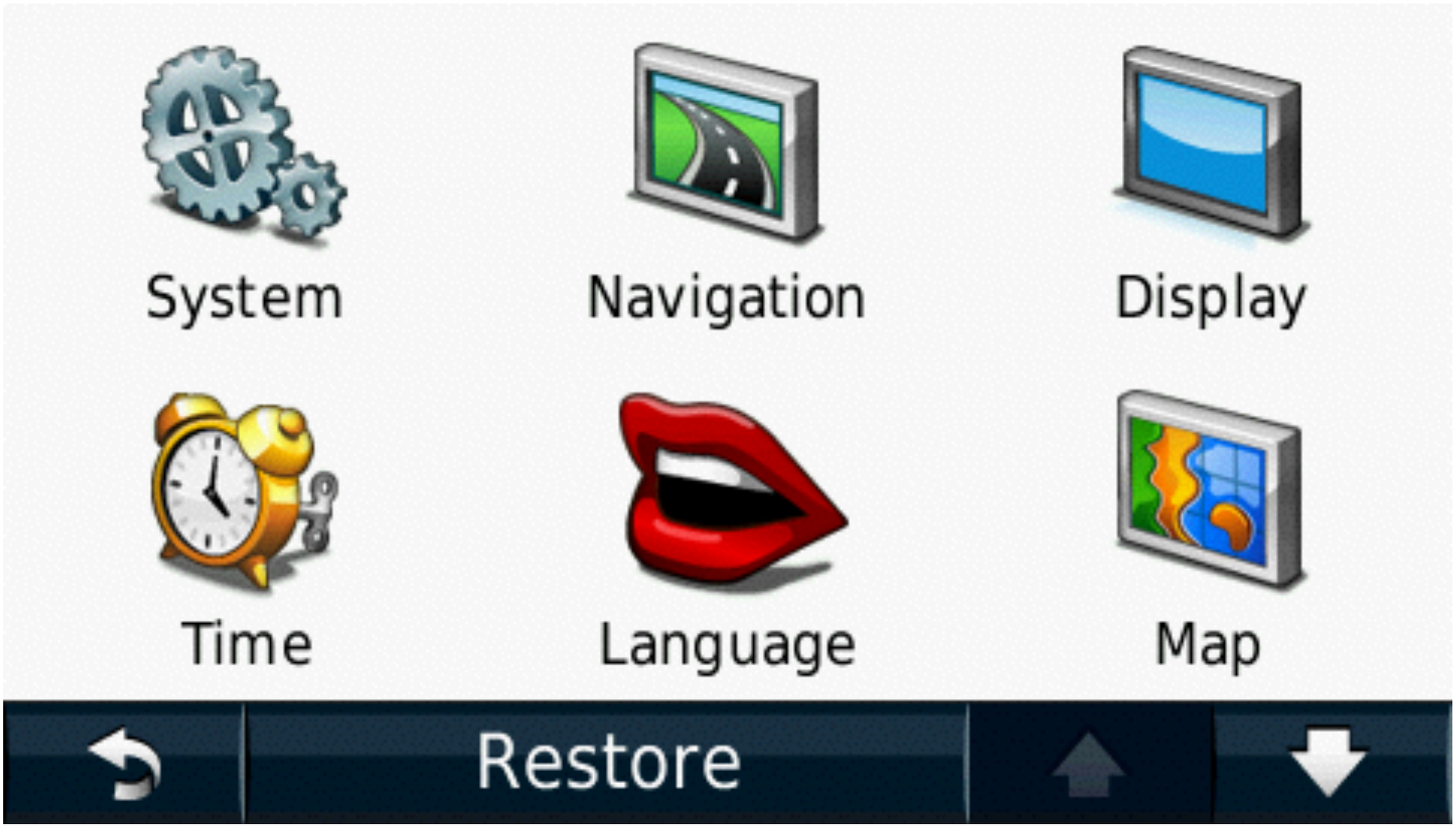
Garmin maps show where you are...

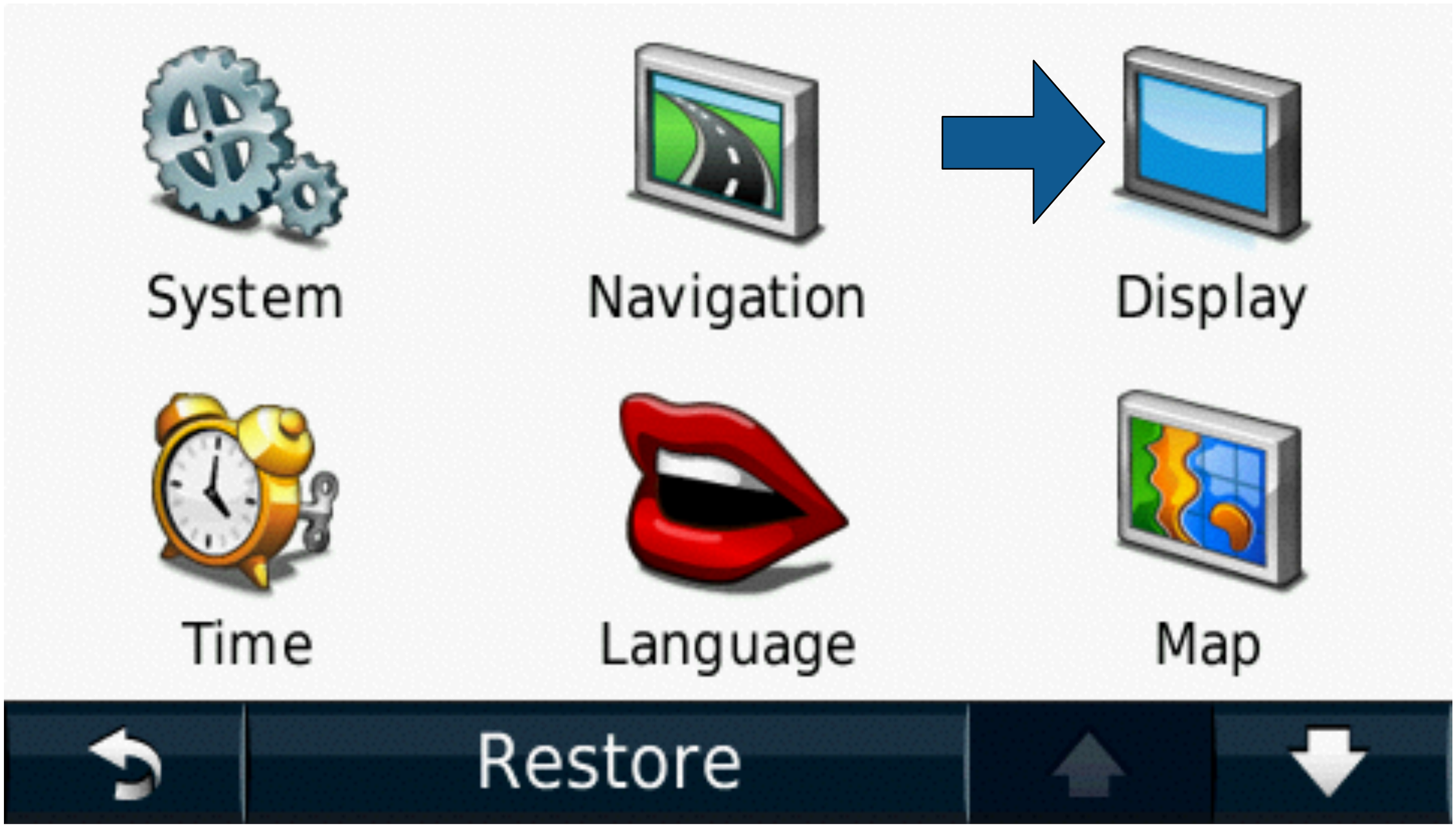


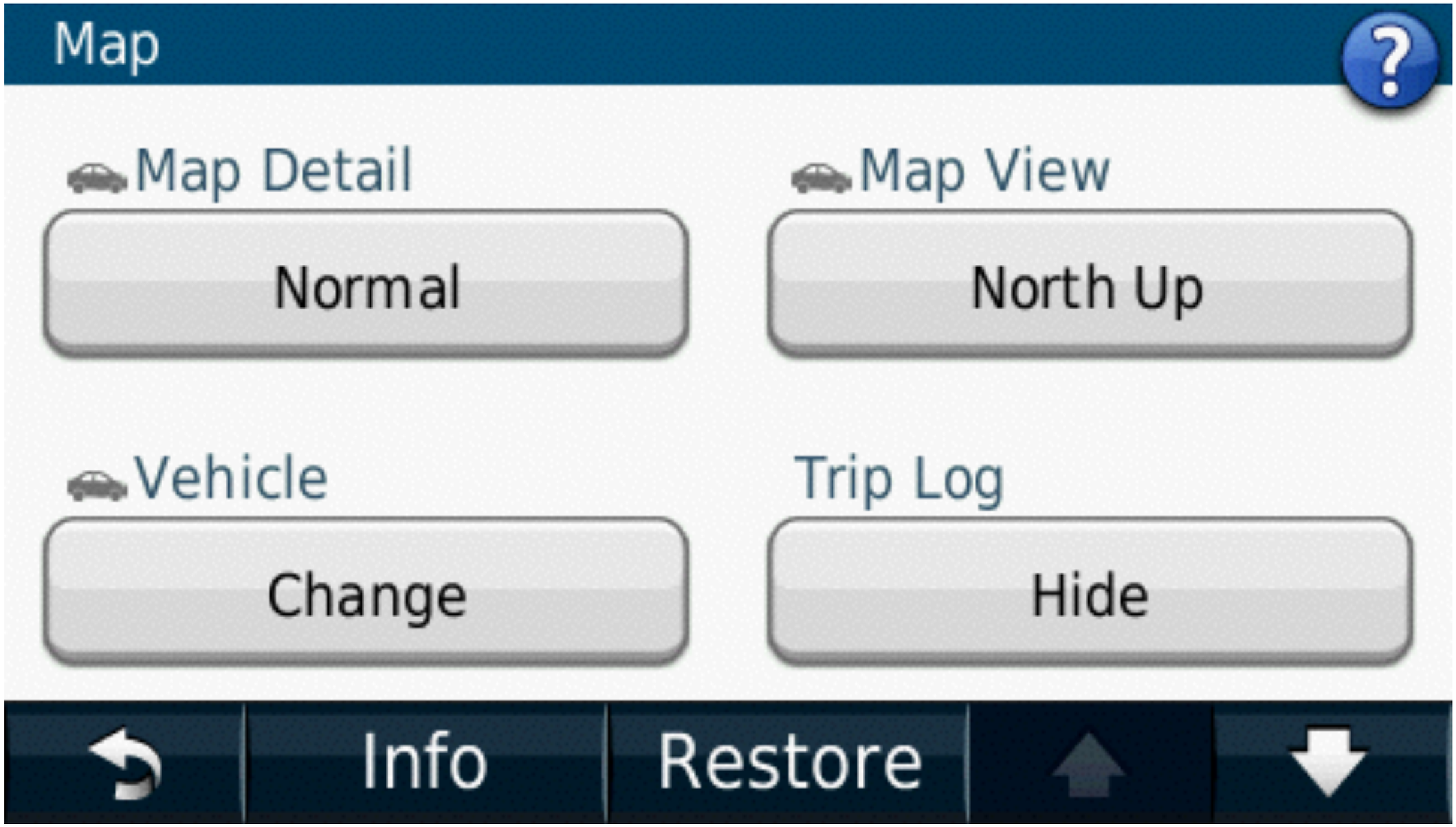
(taken with Garmin's screen capture)

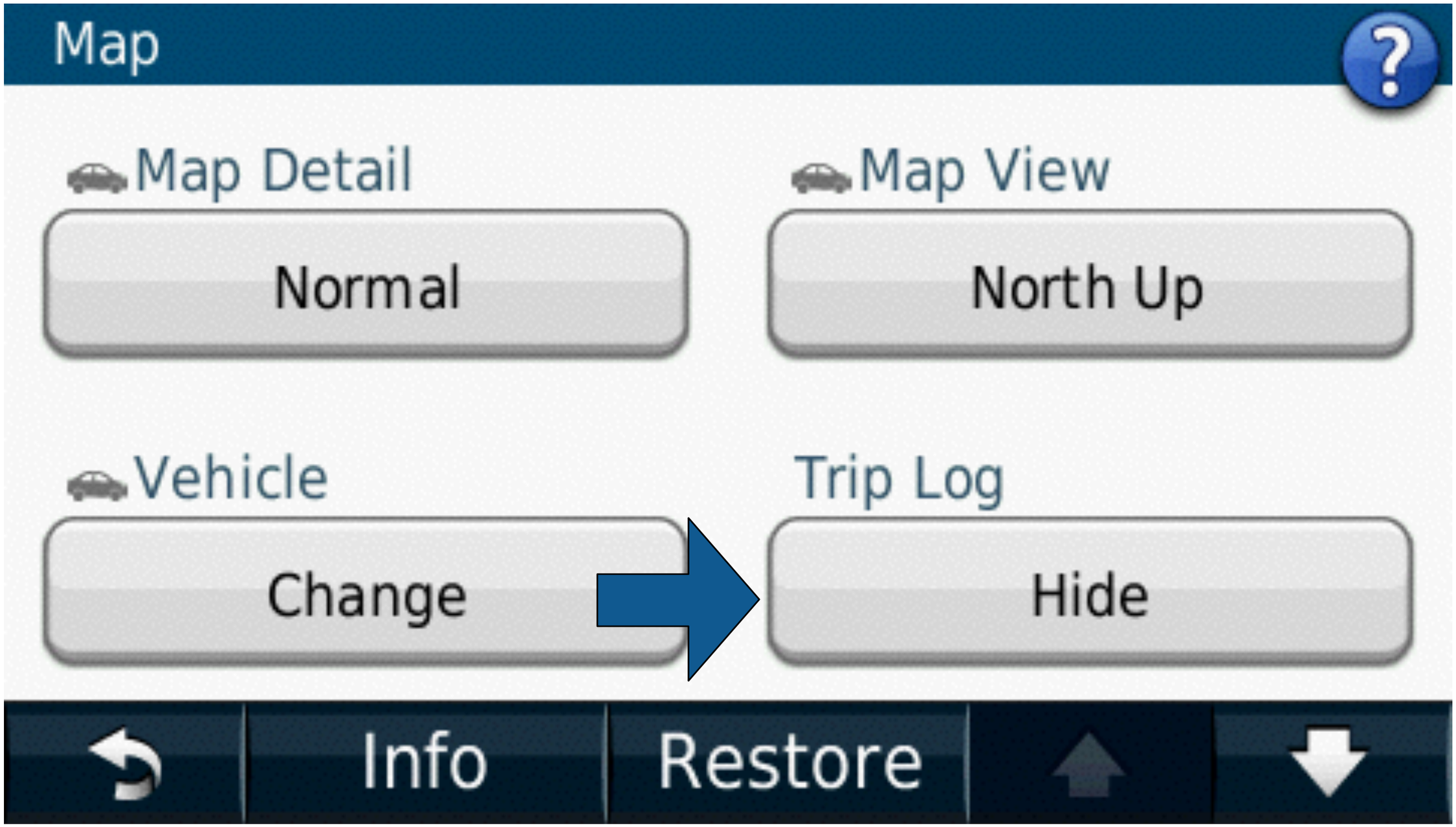

Settings
Where Am I?
Help
ecoRoute™
Picture Viewer
My Data

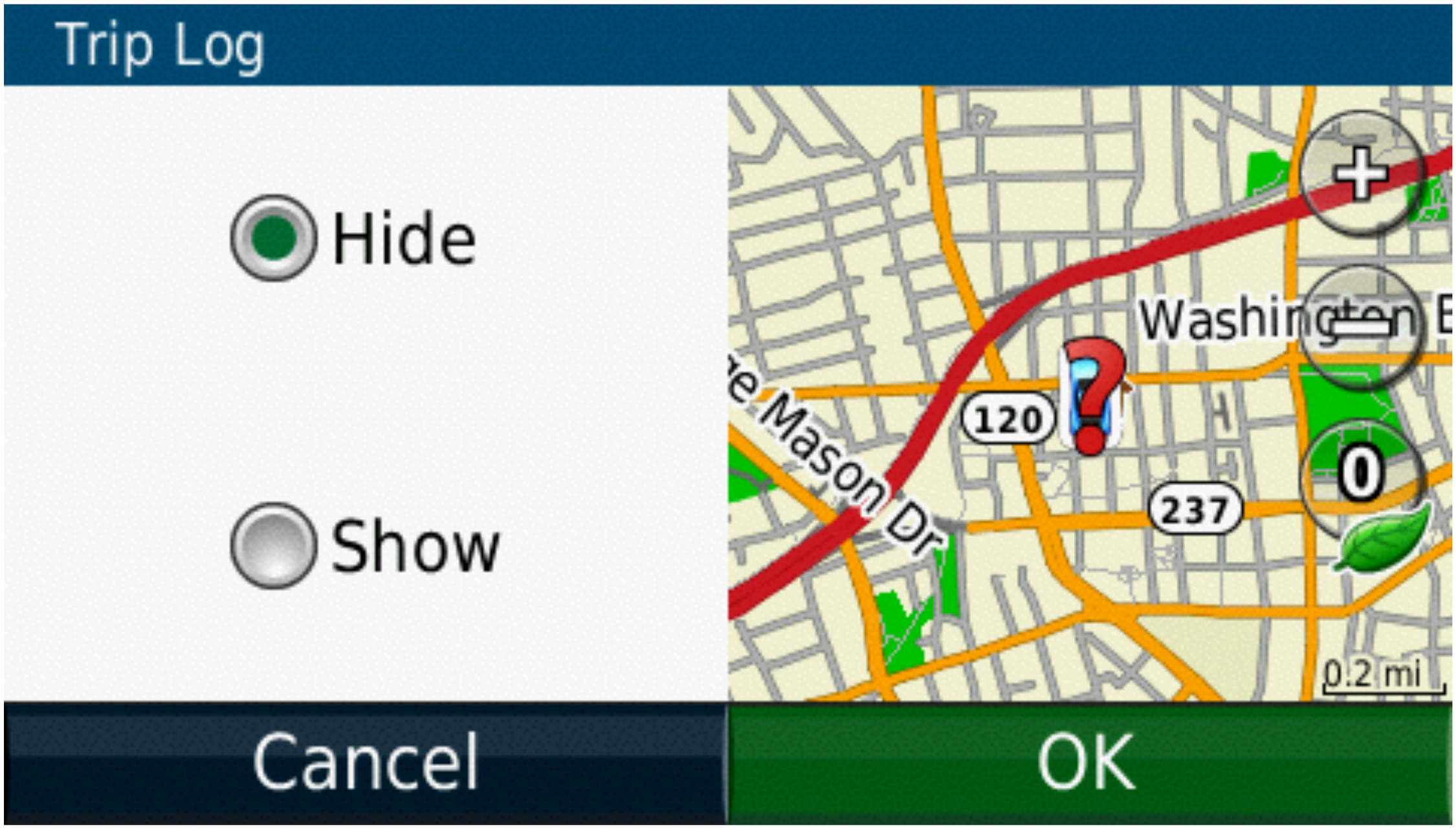

Settings
Where Am I?
Help
ecoRoute™
Picture Viewer
My Data

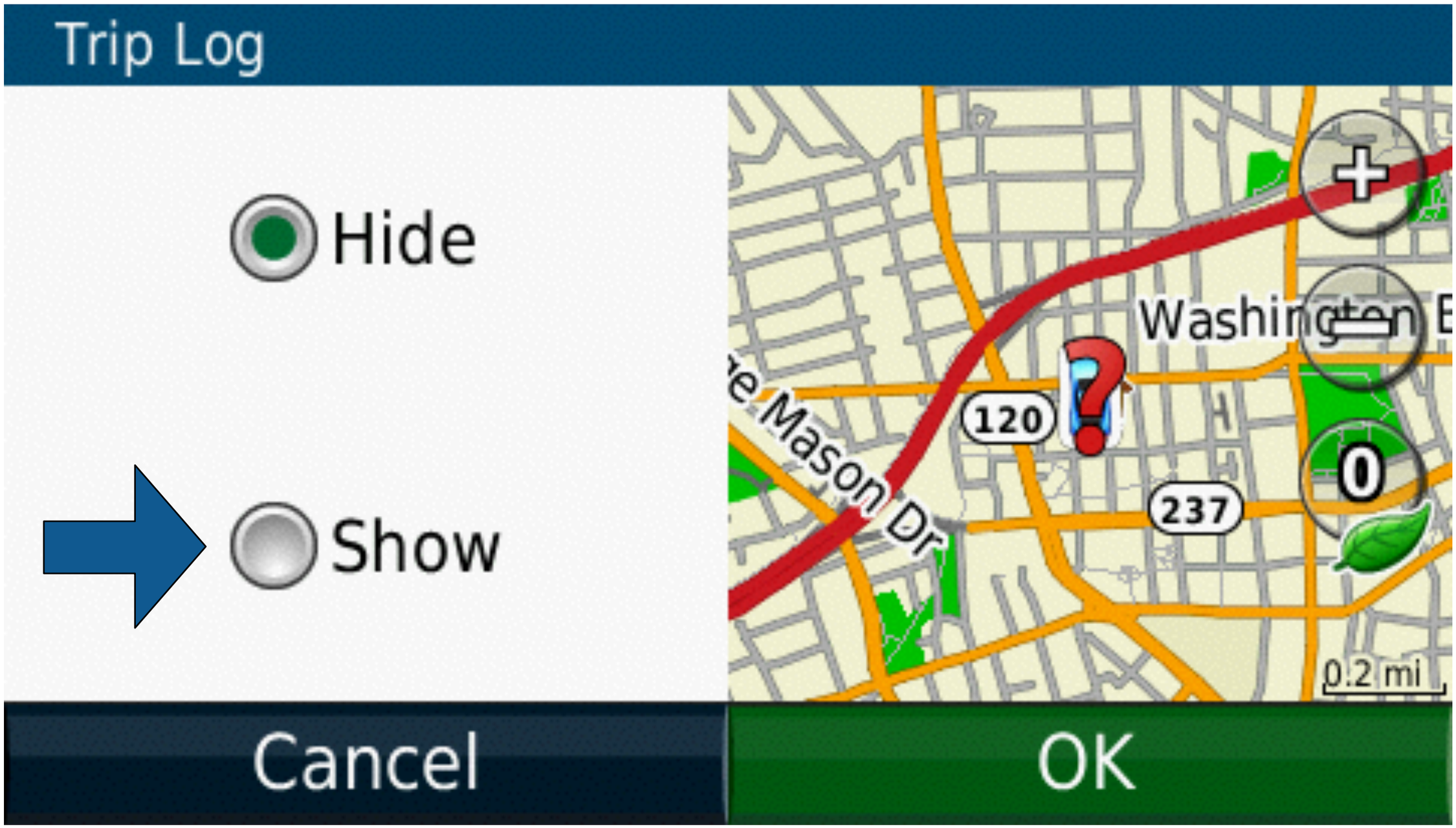

System
Navigation
Display
Time
Language
Map
Restore


System
Navigation
Display
Time
Language
Map
Restore

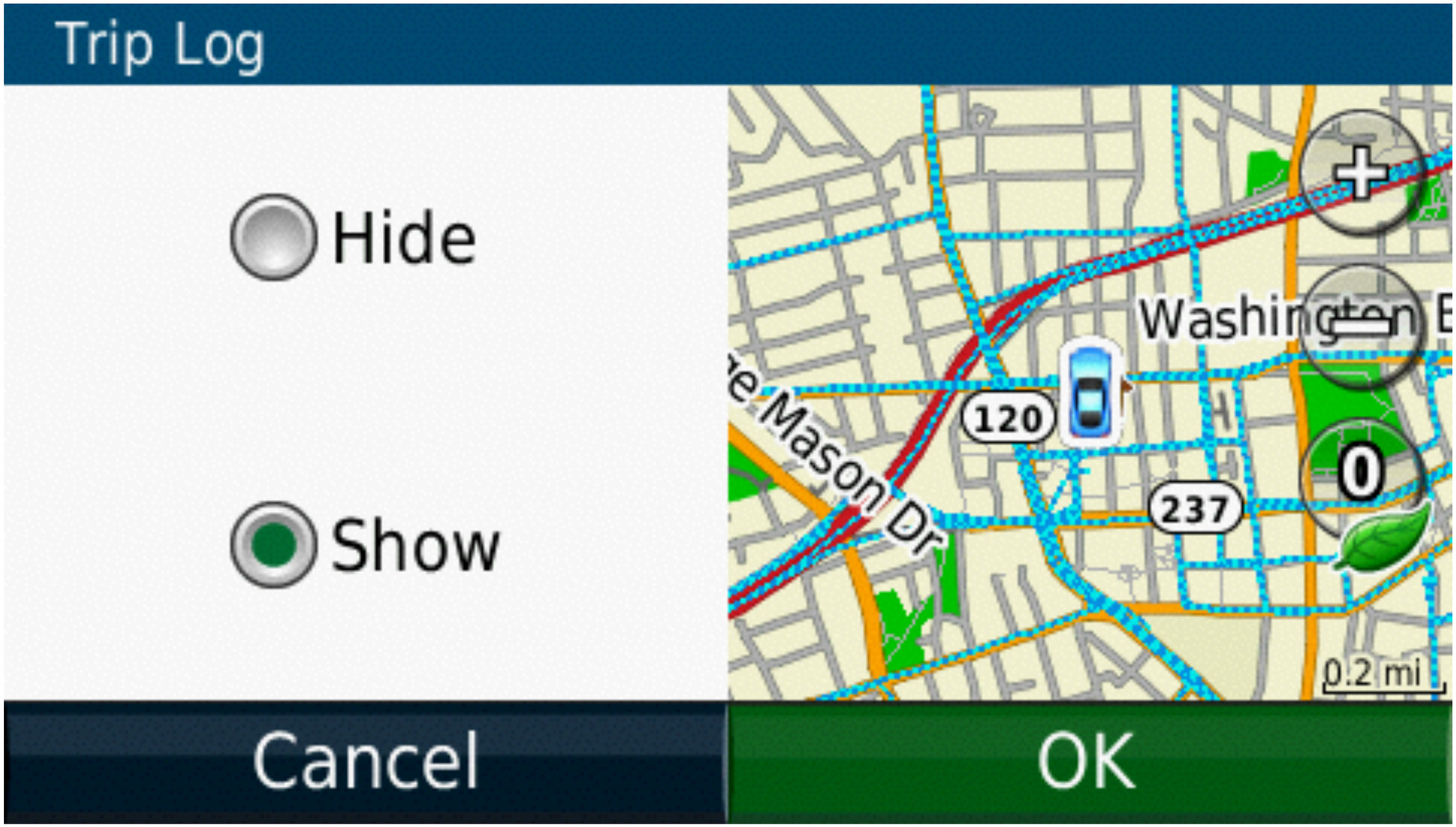

Map
Map Detail
Normal
Map View
North Up
Vehicle
Change
Trip Log
Hide
Info
Restore


Map
Map Detail
Normal
Map View
North Up
Vehicle
Change
Trip Log
Hide
Info
Restore

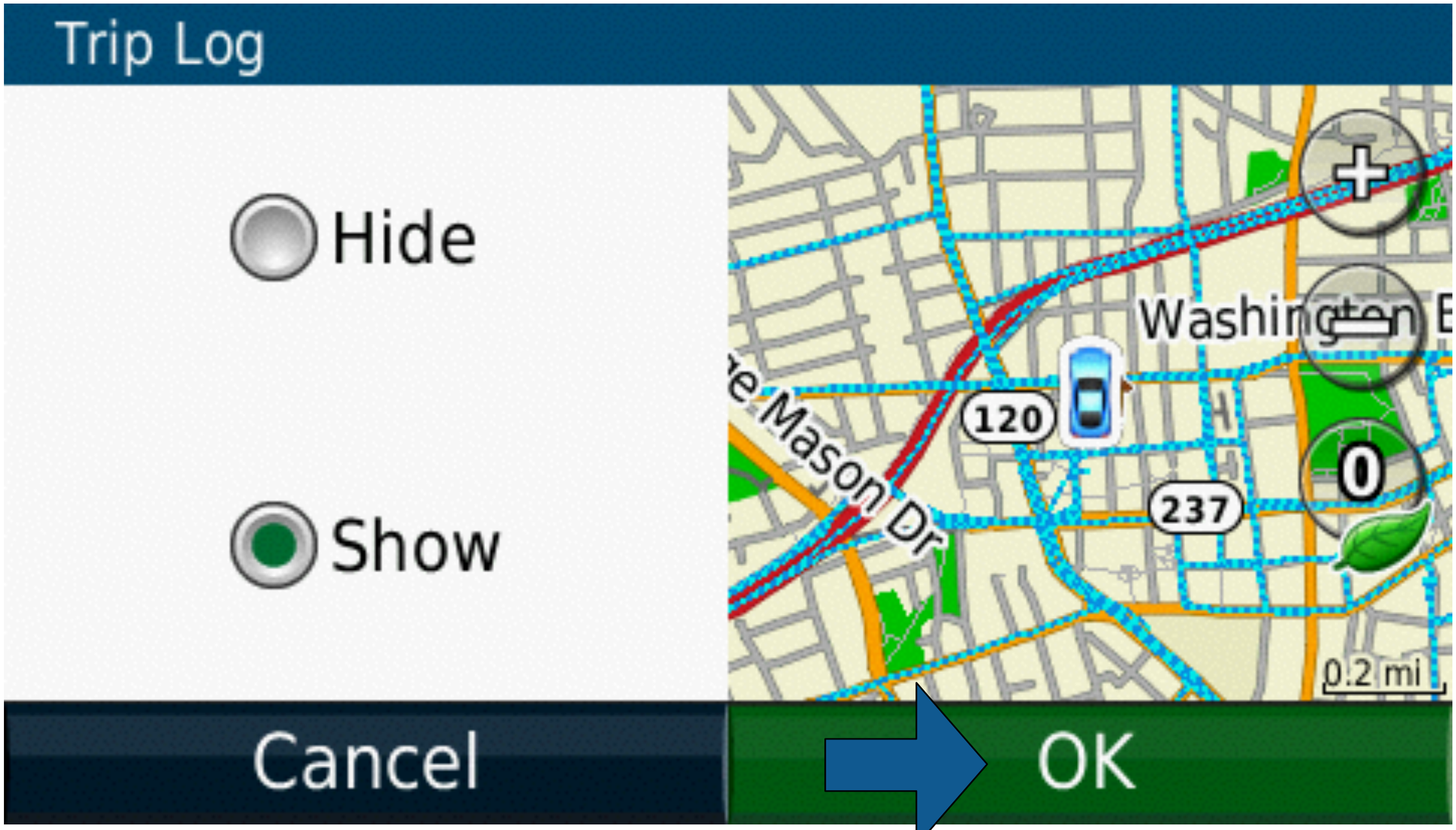

Trip Log
Hide
Show
Washington E
e Mason Dr
120
237
0
0.2 mi
Cancel
OK

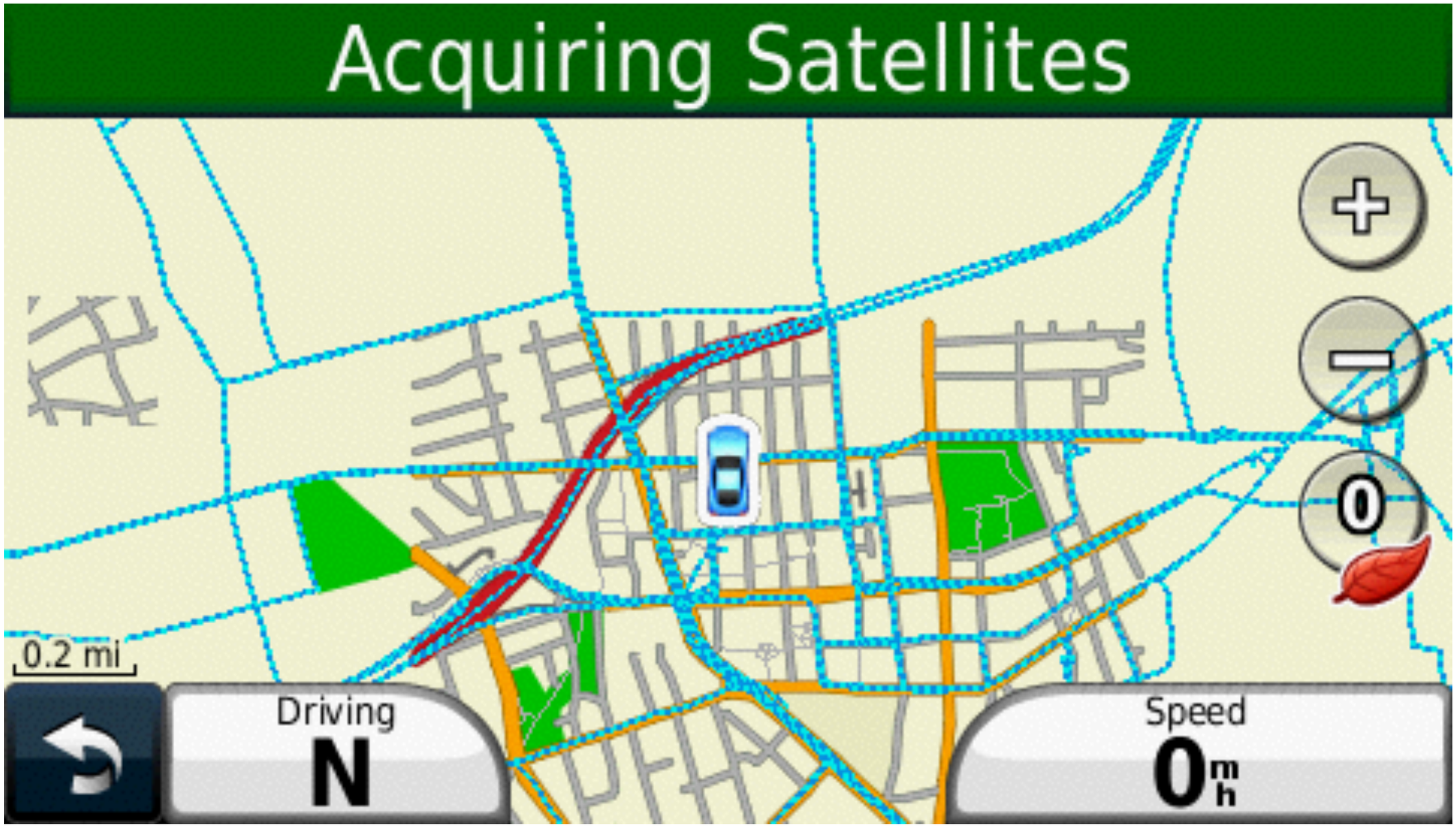

Trip Log
Hide
Show
Washington E
e Mason Dr
120
237
0
0.2 mi
Cancel
OK

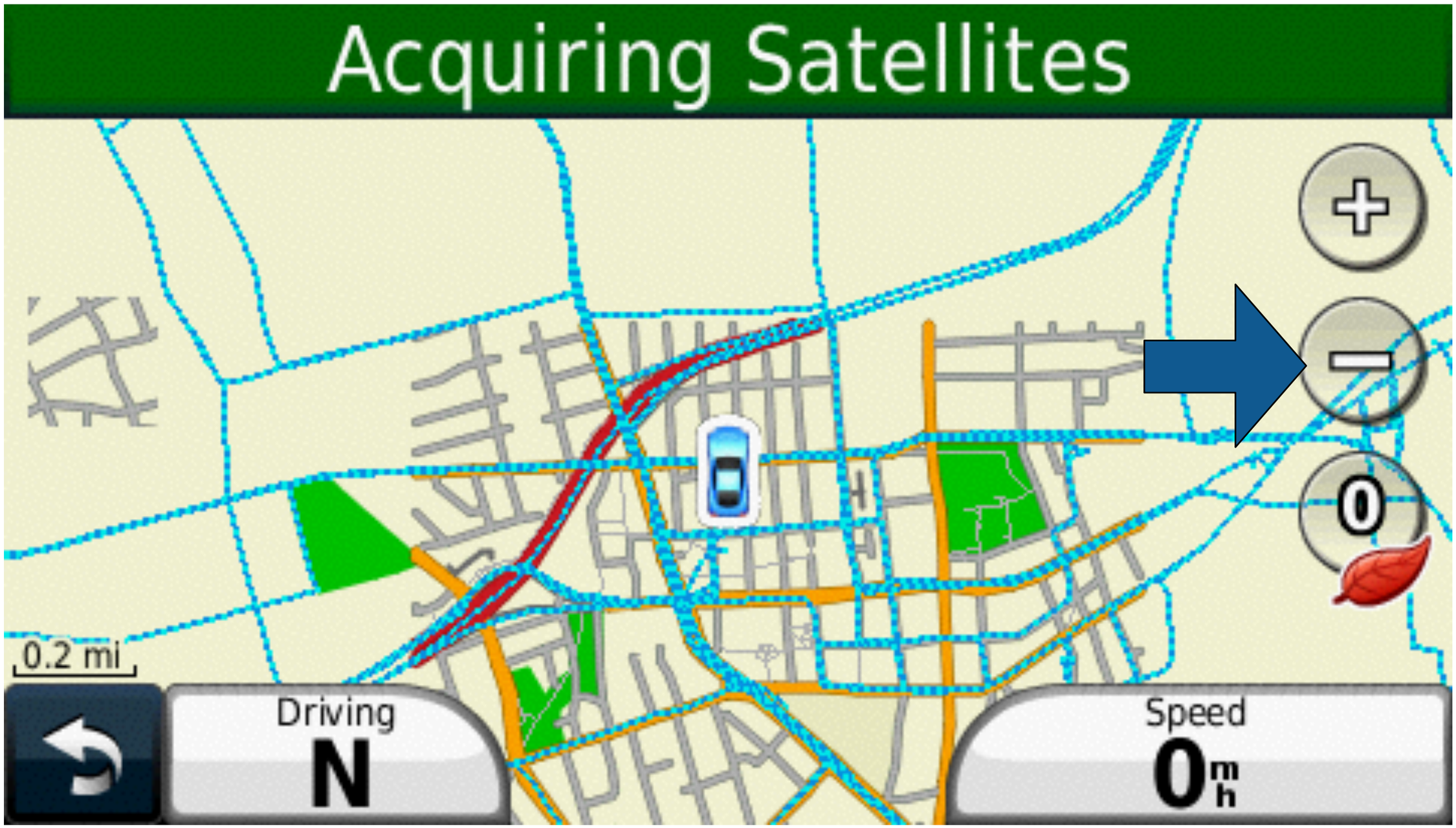
Maps can also show where you have been.

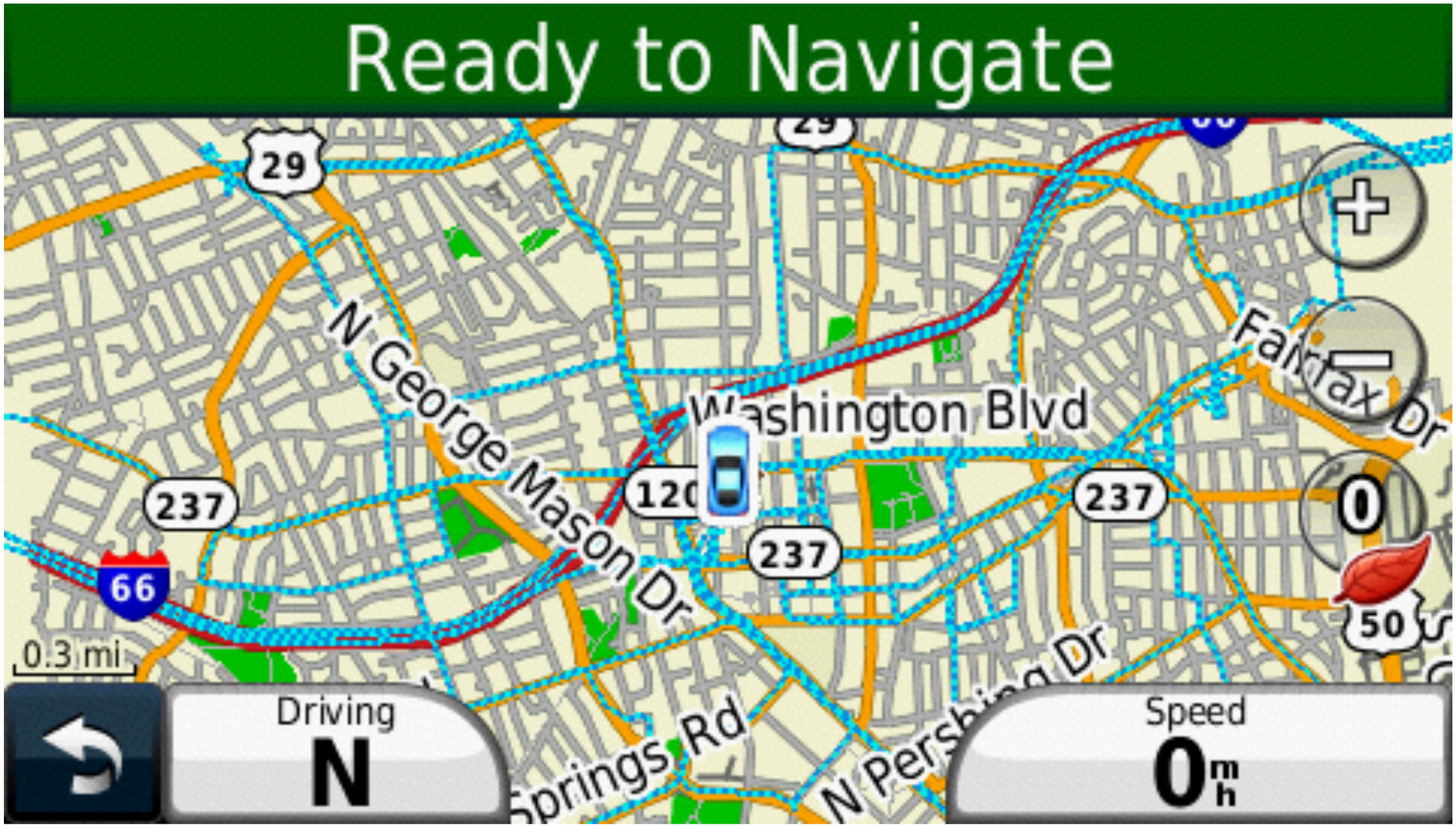


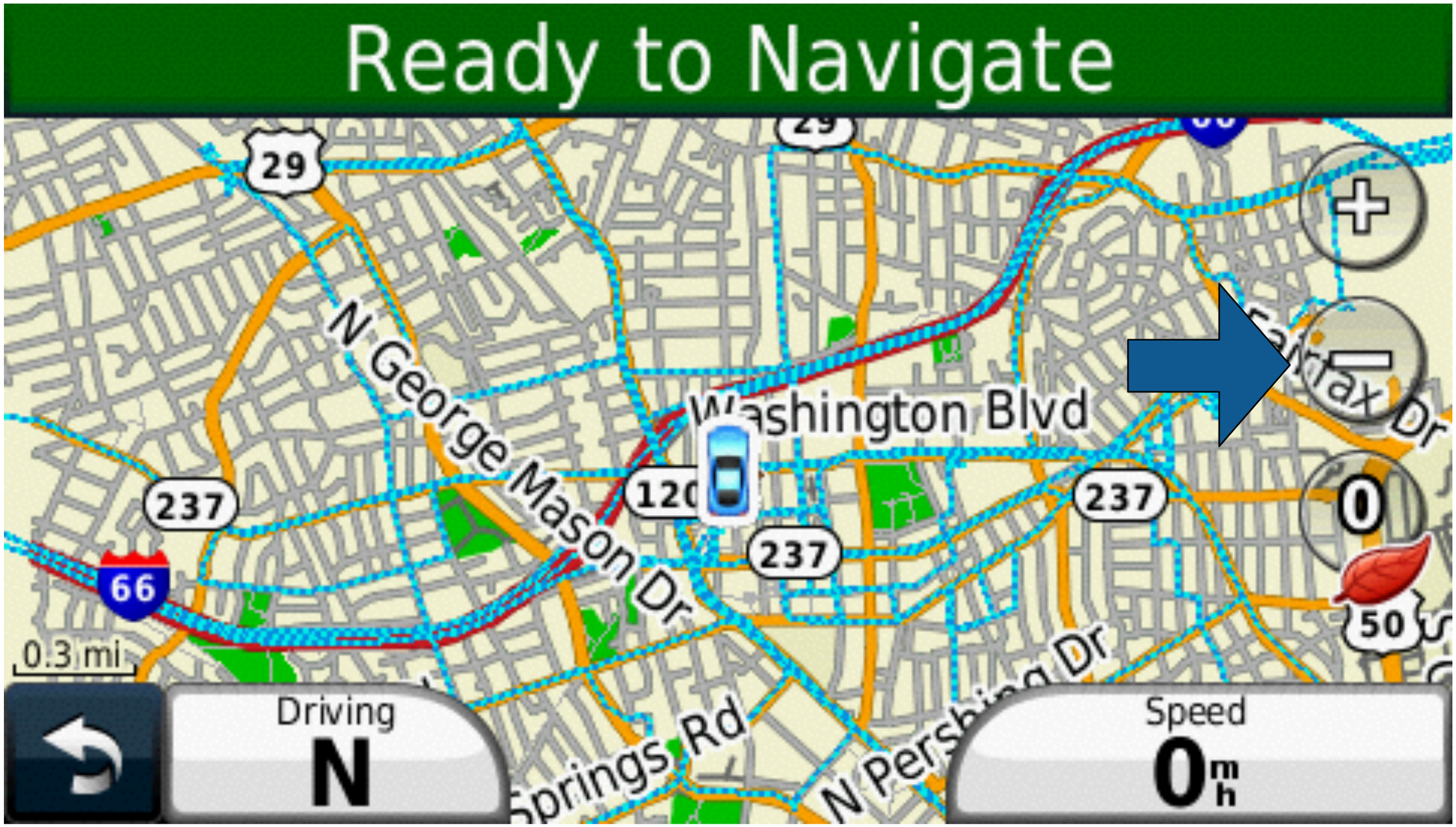
Maps can also show where you have been.

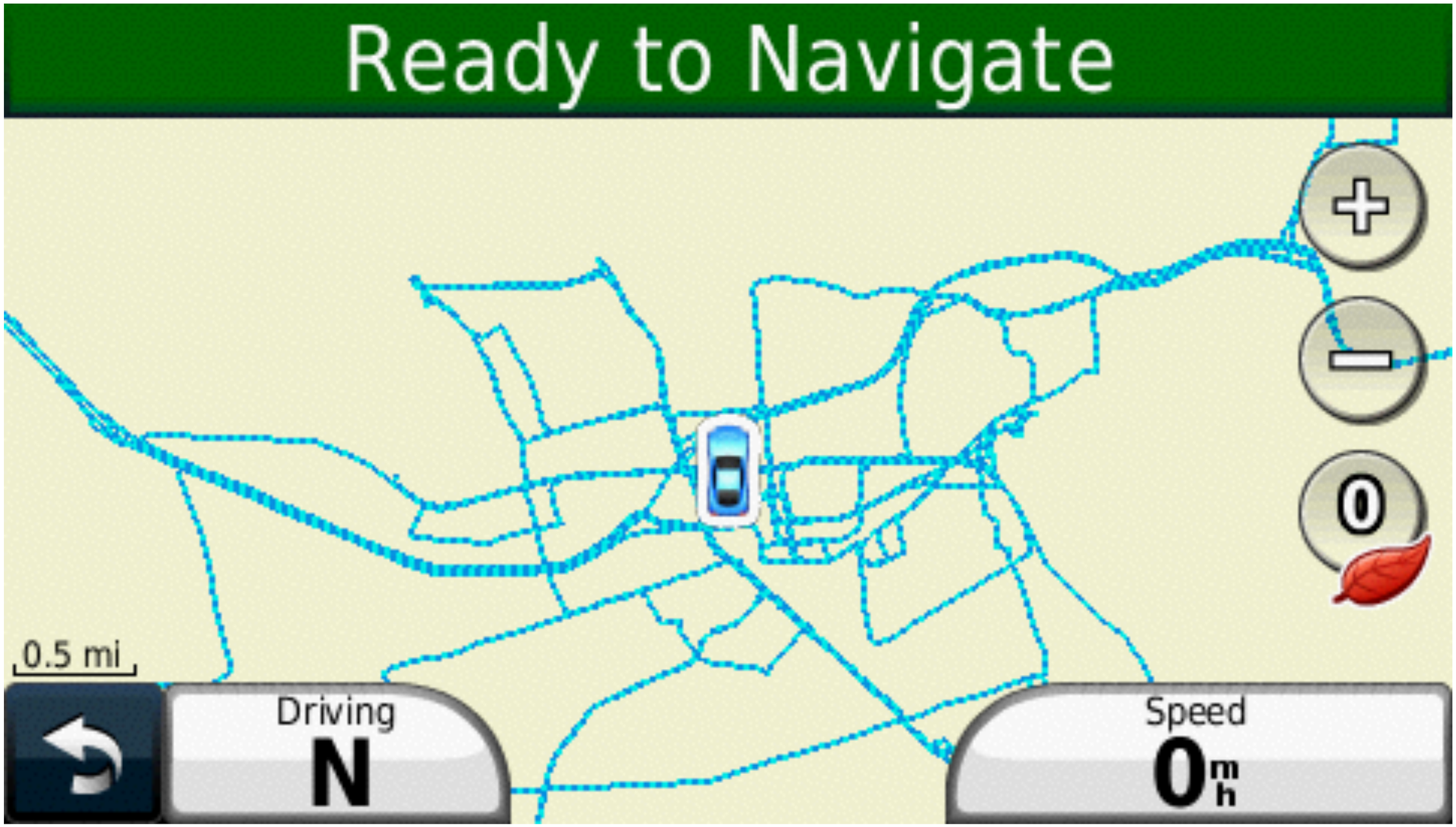


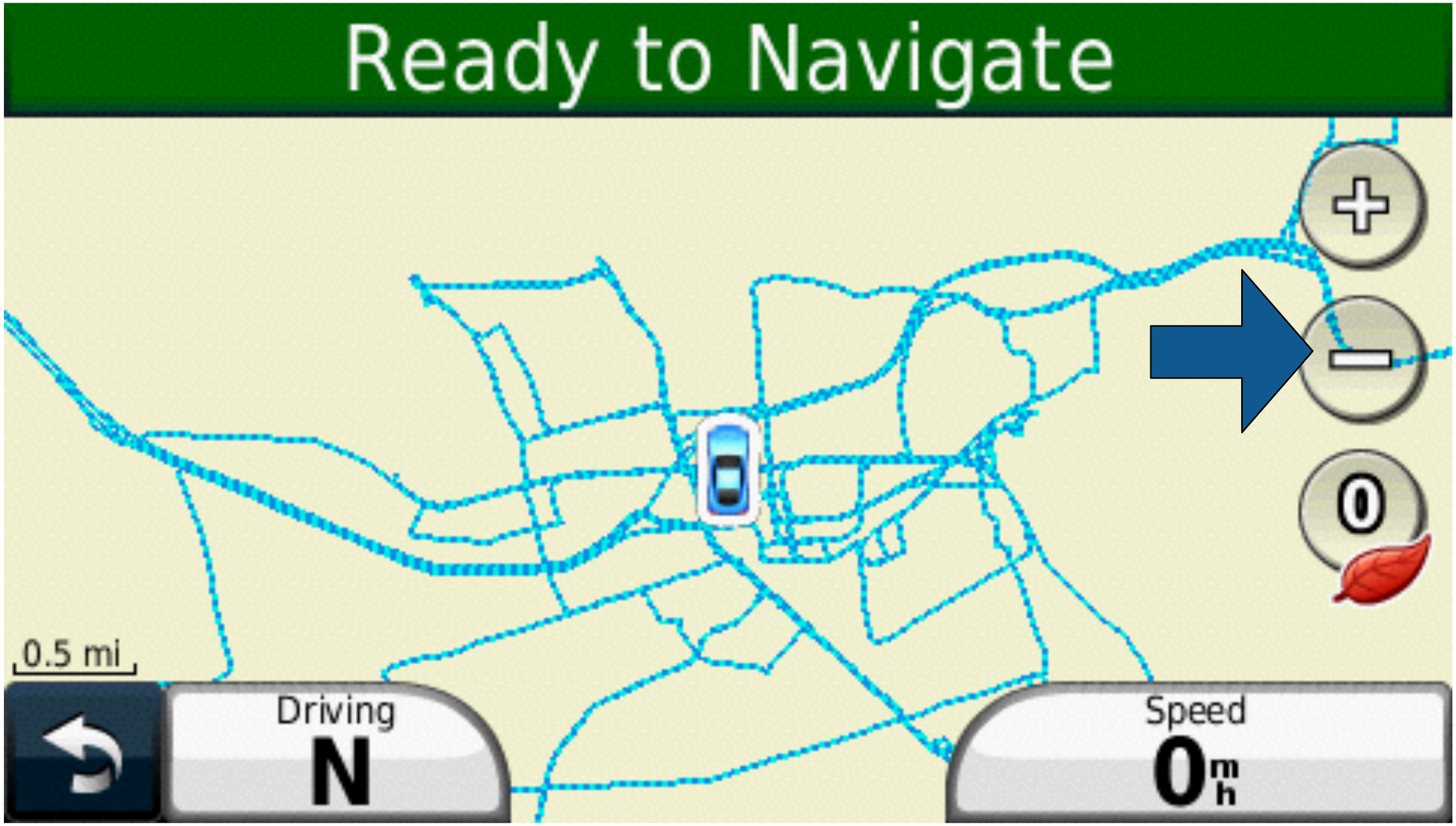

Acquiring Satellites
0.2 mi
Driving
N
Speed
0
m
h
0

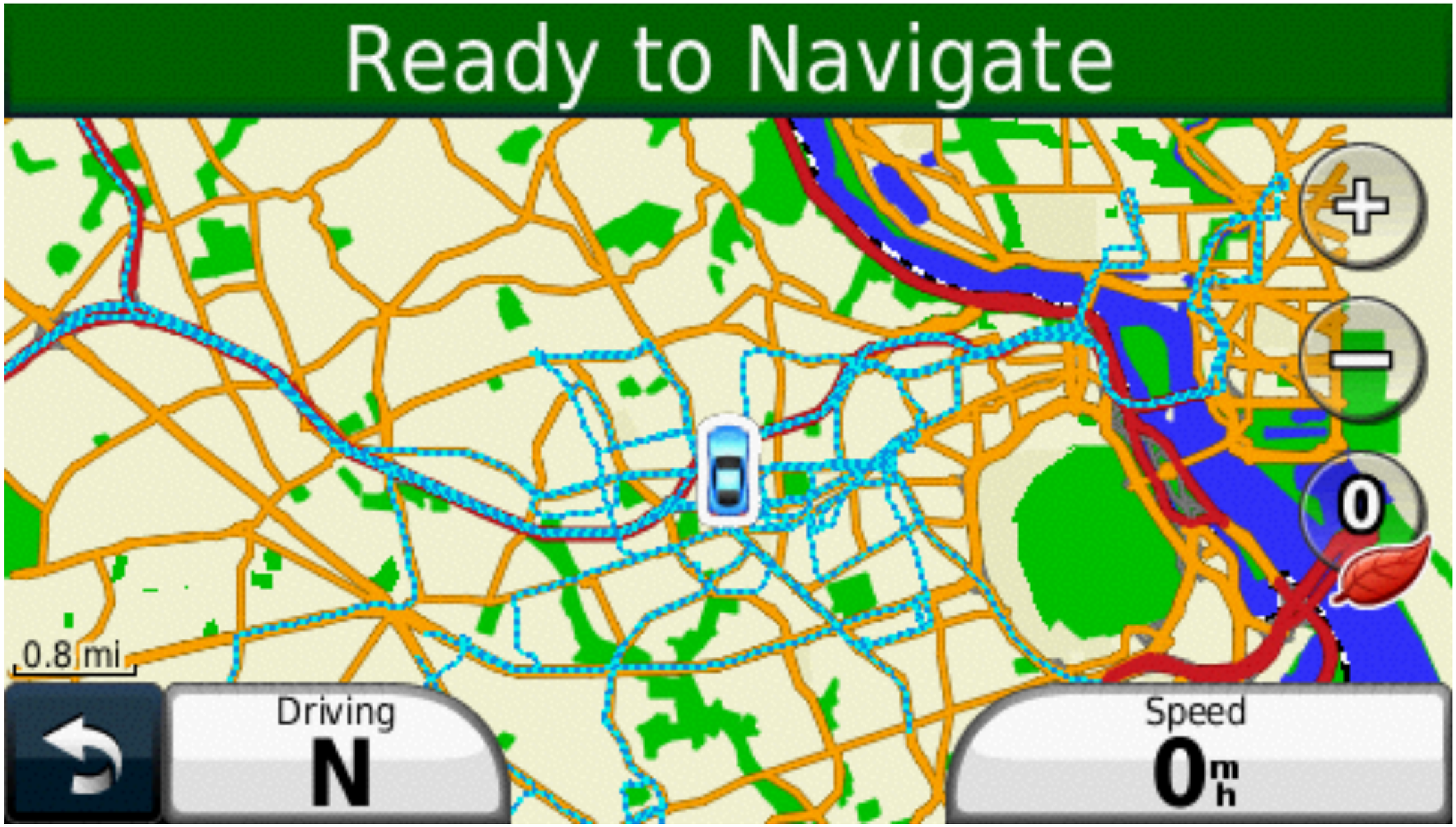

Acquiring Satellites
0.2 mi
Driving
N
Speed
0
m
h
0

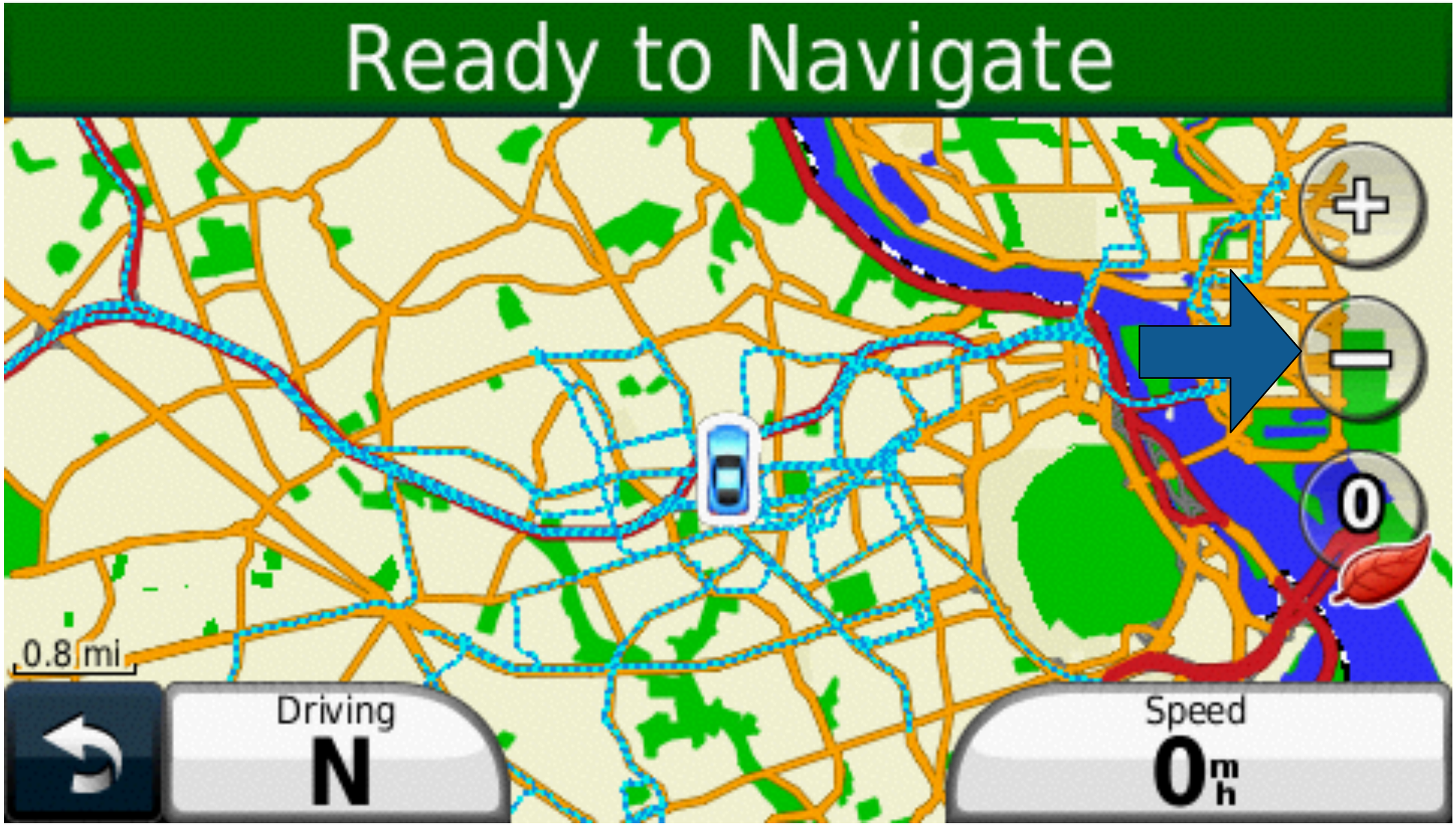

Ready to Navigate
29
29
N George Mason Dr
Washington Blvd
Fairfax Dr
237
120
237
237
66
0
50
0.3 mi
Driving
N
Speed
0
m
h

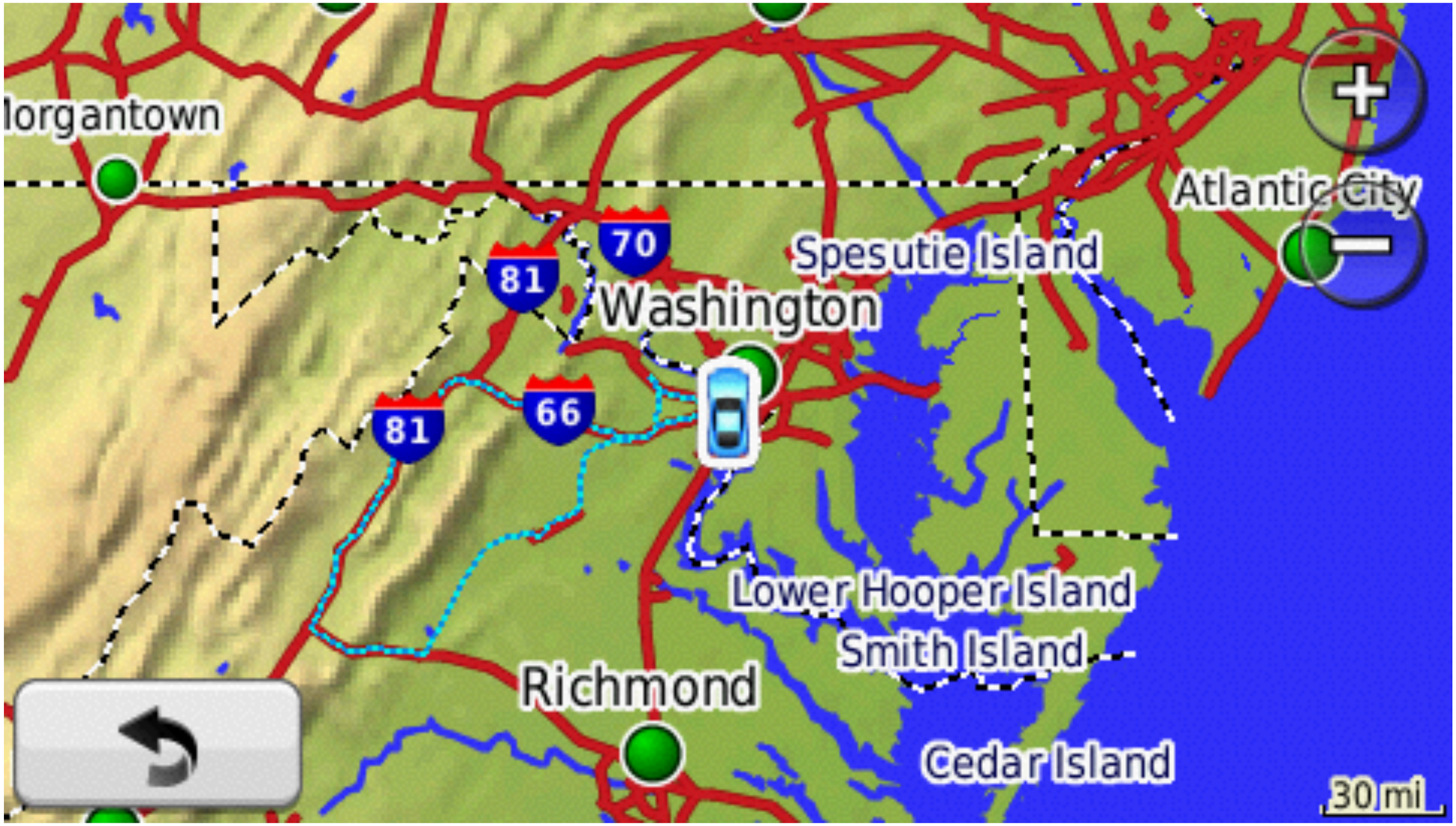

Ready to Navigate
29
29
N George Mason Dr
Washington Blvd
237
120
237
237
66
0.3 mi
Driving
N
Speed
0 m/h
Fairfax Dr
0
50

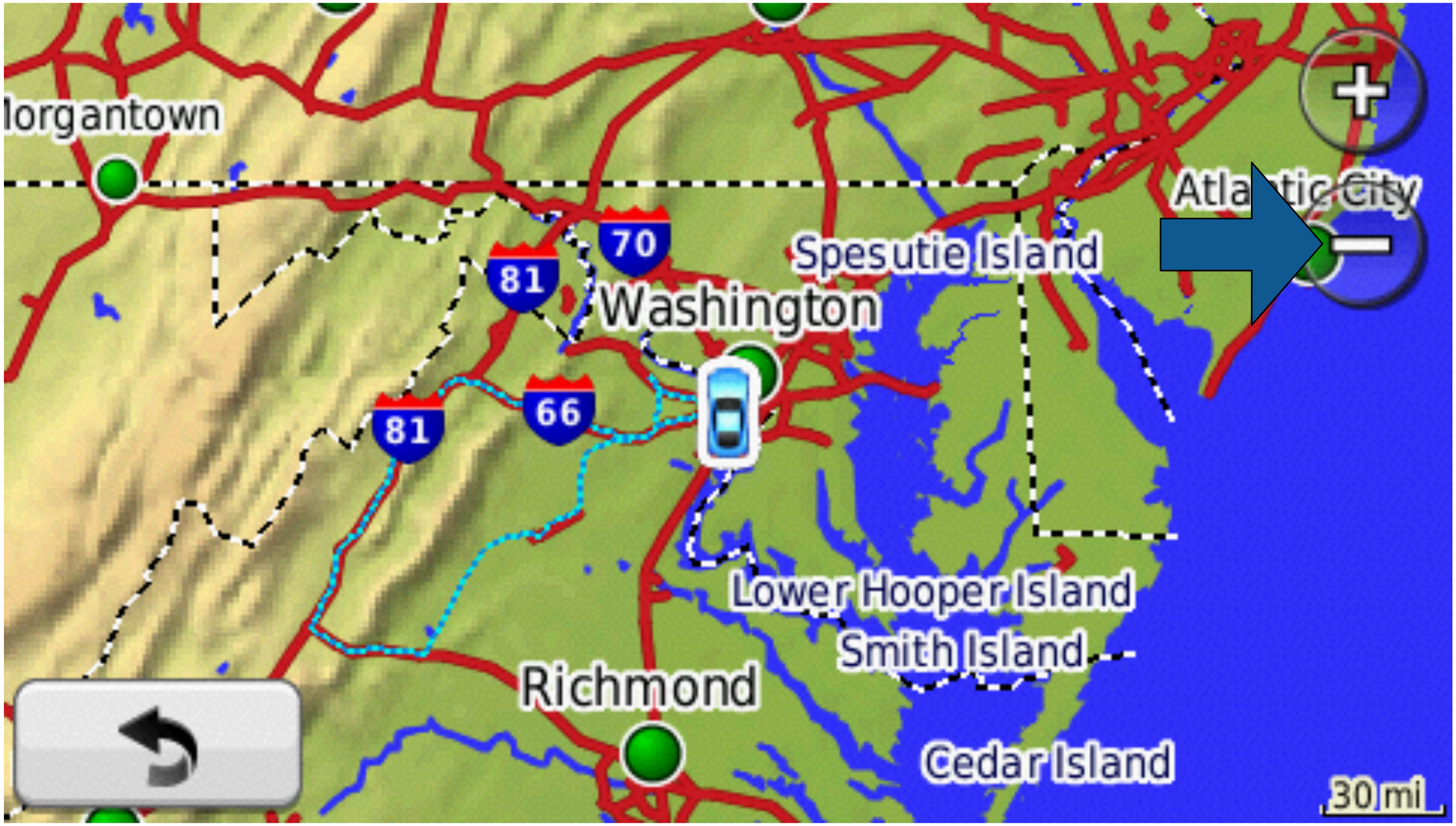

Ready to Navigate
0.5 mi
Driving
N
Speed
0
m
h


Ready to Navigate
0.5 mi
Driving
N
Speed
0
m
h
0


Ready to Navigate
0.8 mi
Driving
N
Speed
0
m
h


Ready to Navigate
0.8 mi
Driving
N
Speed
0
m
h

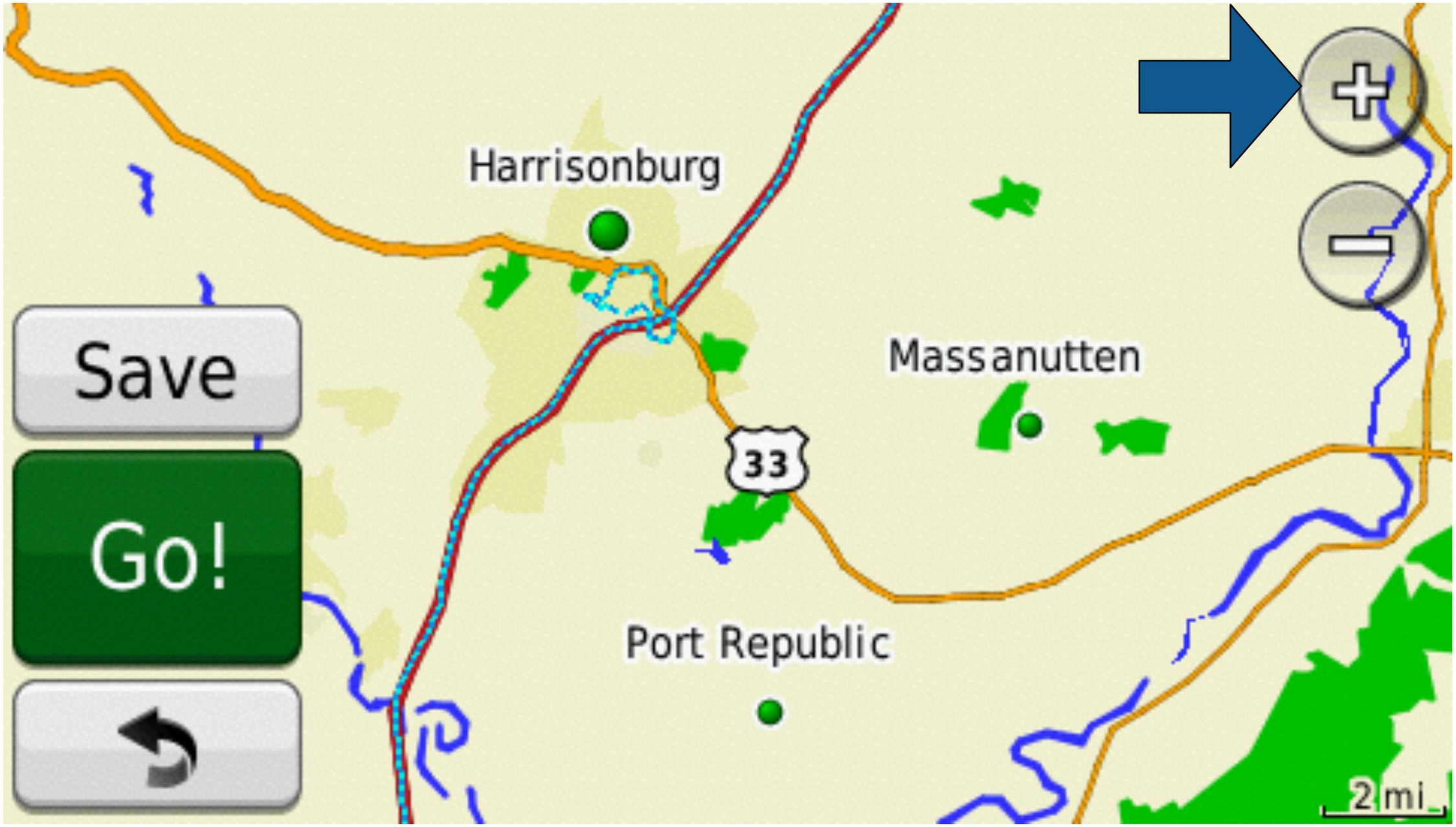

Morgantown
70
81
Washington
Spesutie Island
Atlantic City
81
66
Lower Hooper Island
Smith Island
Richmond
Cedar Island
30 mi

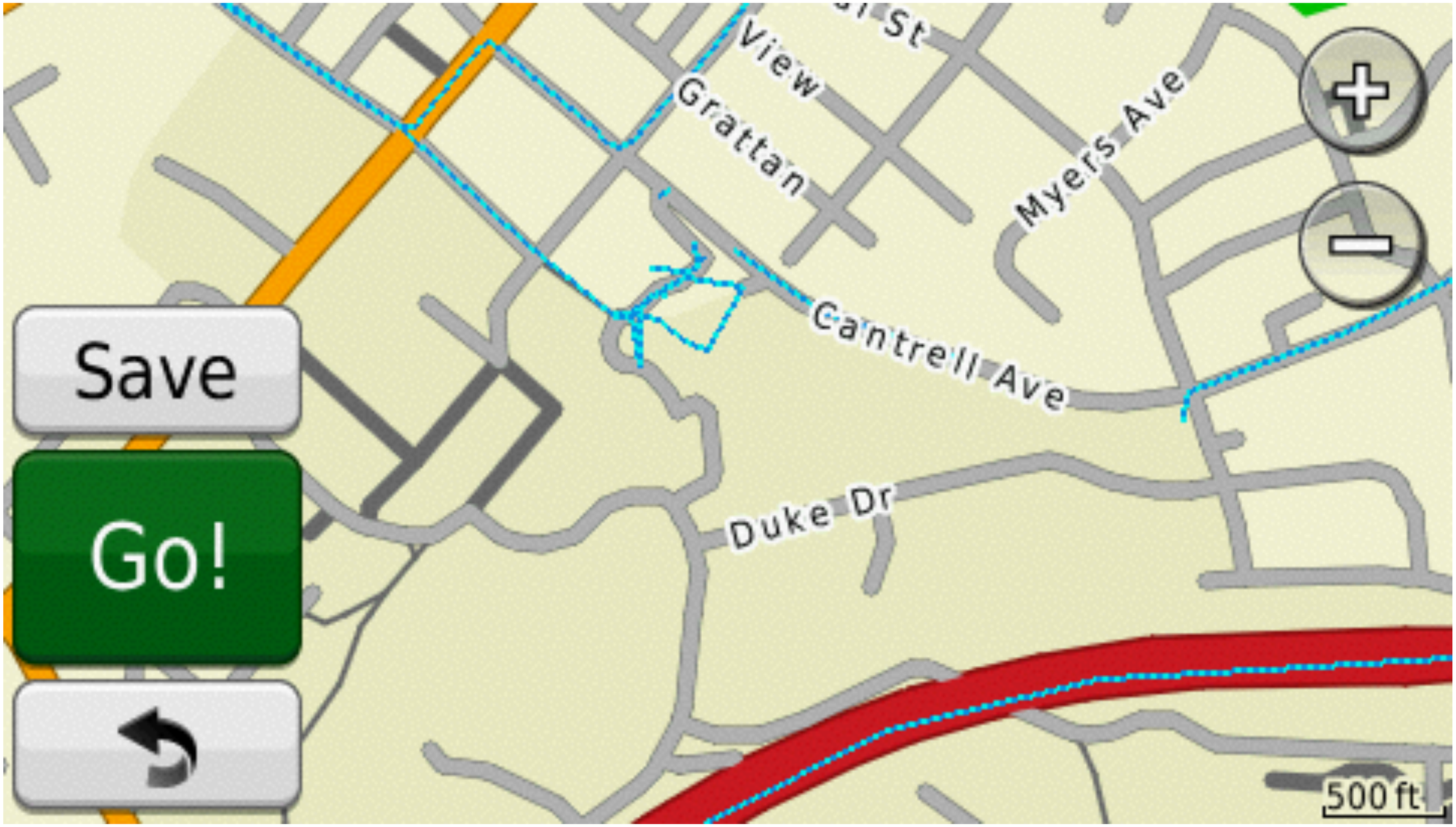

Morgantown
70
81
Washington
Spesutie Island
Atlantic City
81
66
Lower Hooper Island
Smith Island
Richmond
Cedar Island
30 mi

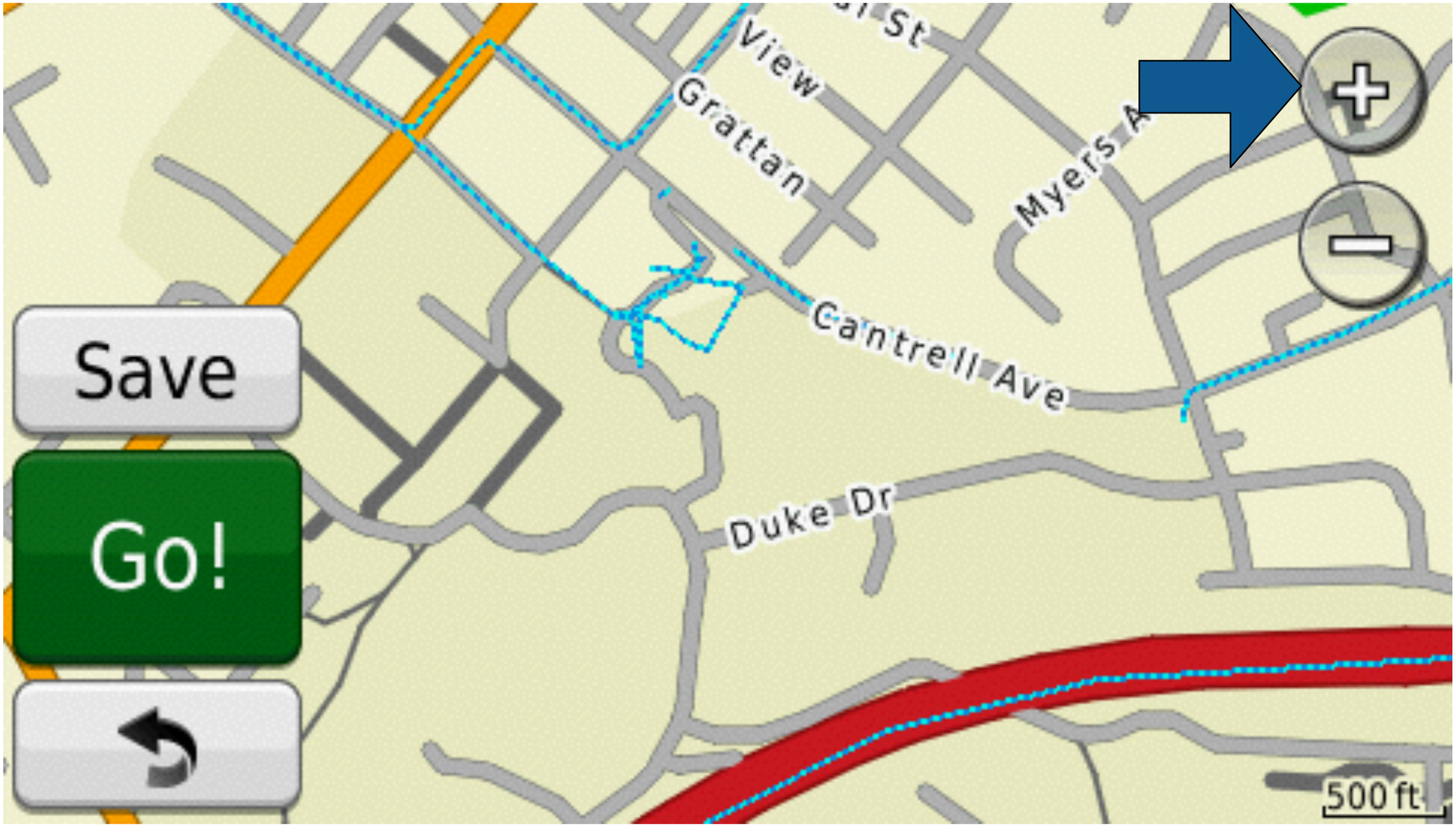

Harrisonburg
Culpeper
Staunton
Charlottesville
8 mi

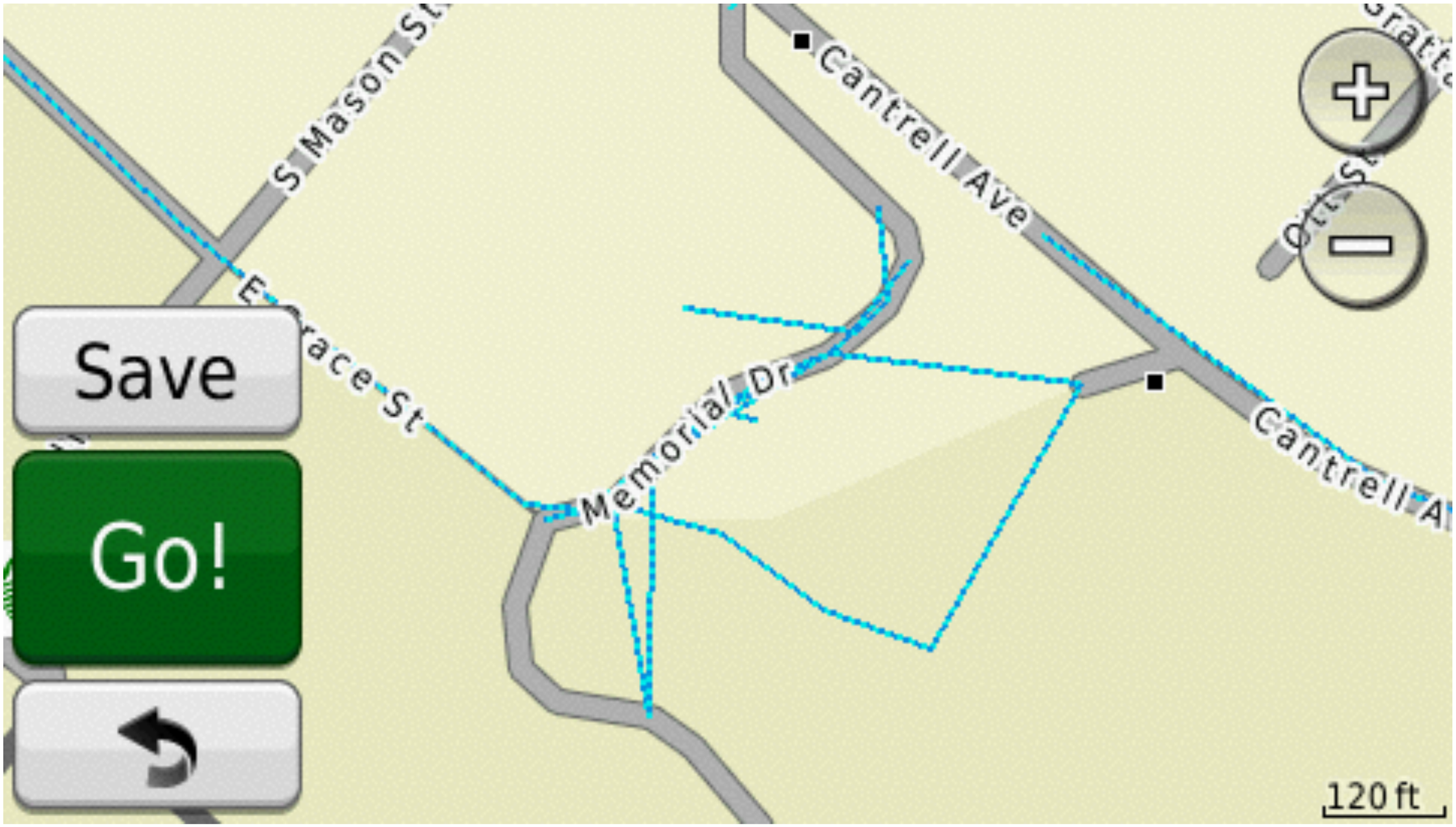

Harrisonburg
Culpeper
Staunton
Charlottesville
8 mi

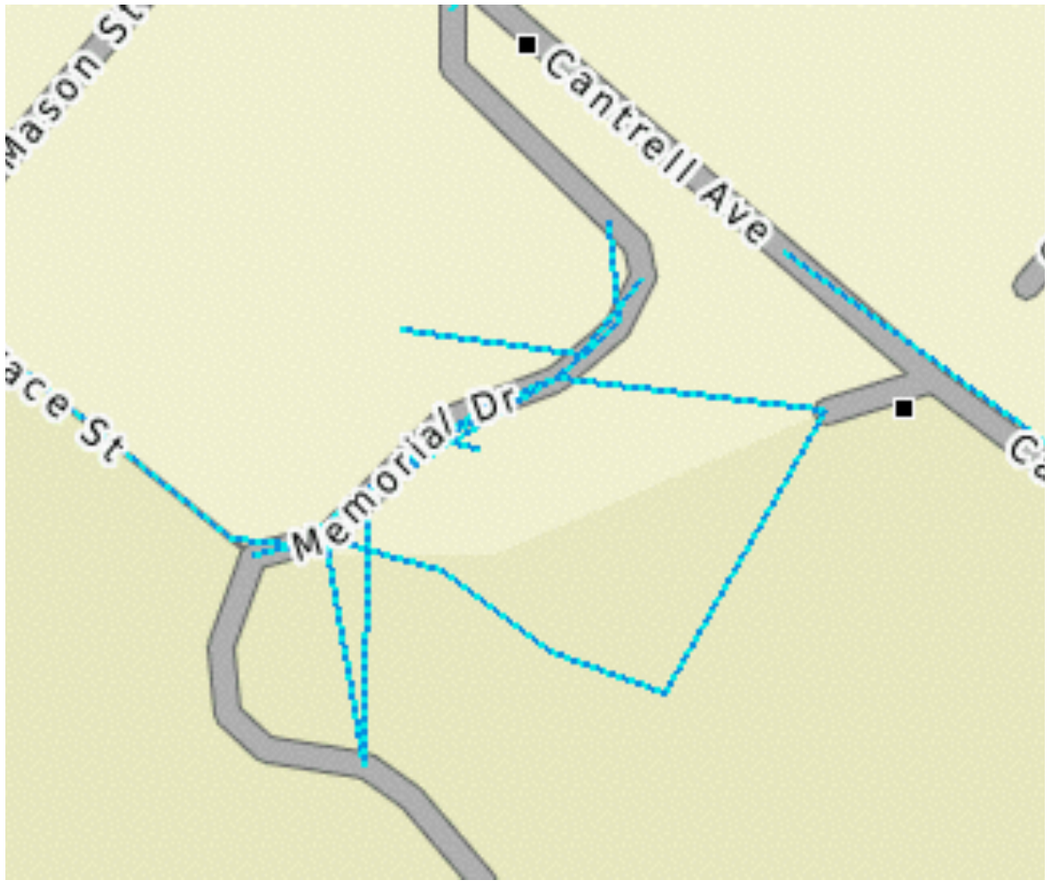

Harrisonburg
Massanutten
33
Port Republic
2 mi
Save
Go!


Harrisonburg
Massanutten
33
Port Republic
2 mi
Save
Go!


View
Grattan
Myers Ave
Cantrell Ave
Duke Dr
Save
Go!
500 ft

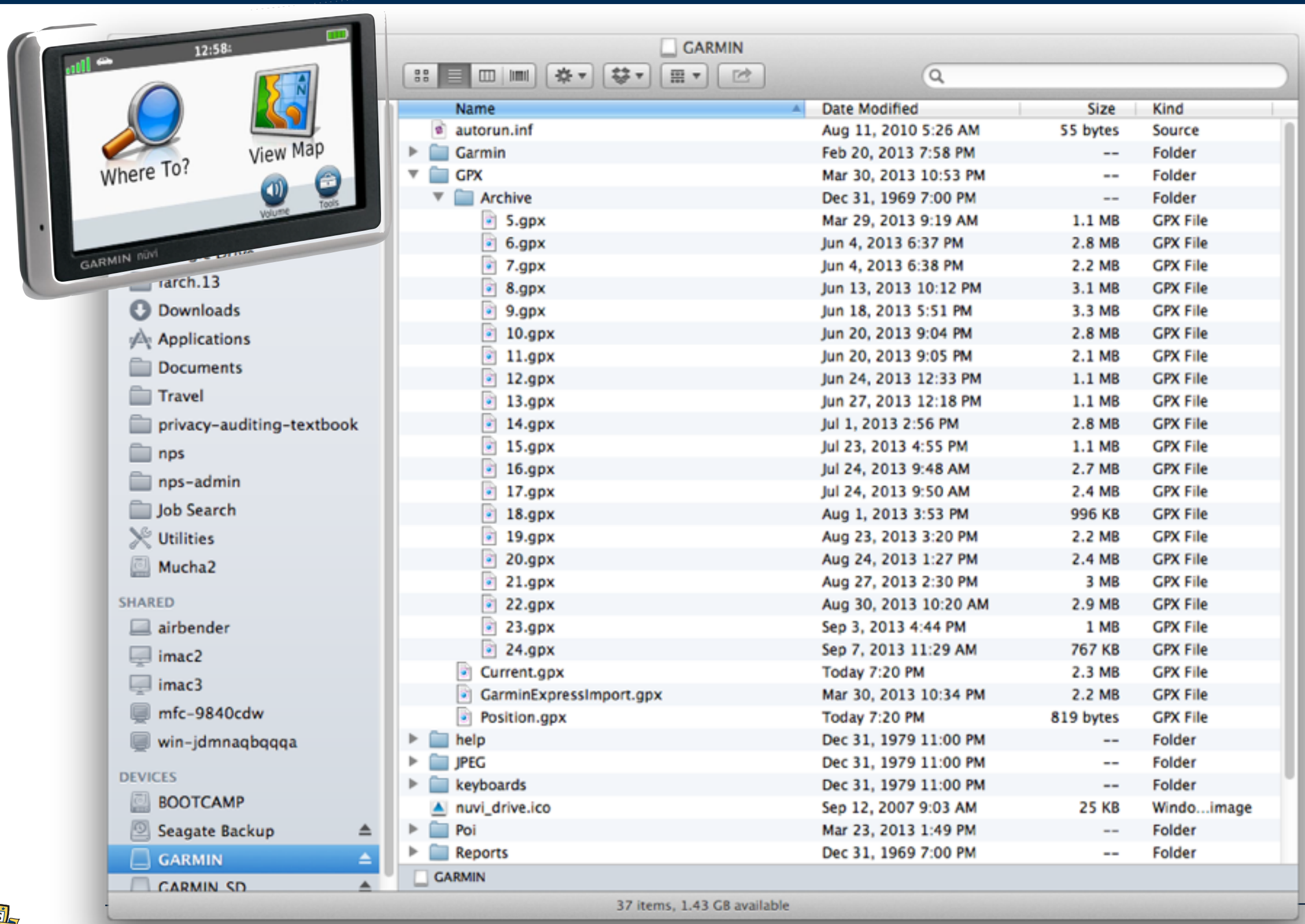

View
Grattan
Myers A
Cantrell Ave
Duke Dr
Save
Go!
500 ft


S Mason St
Cantrell Ave
Memorial Dr
Cantrell Ave
Save
Go!
120 ft

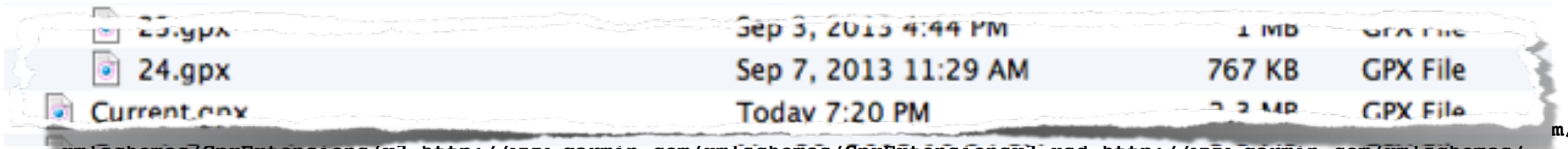

Cantrell Ave
Memorial Dr


S Mason St
Cantrell Ave
E Grace St

The Garmin GPS appears as an external USB storage device with directories and files.



The file 24.gpx contains track information in XML format.




```
m/
xmlschemas/GpxExtensions/v3 http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.garmin.com/xmlschemas/
TrackPointExtension/v2 http://www.garmin.com/xmlschemas/TrackPointExtensionv2.xsd"><metadata><link href="http://
www.garmin.com"><text>Garmin International</text></link><time>2013-09-03T20:44:09Z</time>
</metadata><trk><name>ACTIVE LOG: 23 AUG 2013 10:47</name><trkseg><trkpt lat="38.885255" lon="-77.114185"><ele>99.63</
ele><time>2013-08-23T14:47:26Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:course>0.00</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt lat="38.885206" lon="-77.114113"><ele>89.54</
ele><time>2013-08-23T14:47:49Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>1.37</
gpxtpx:speed><gpxtpx:course>0.00</gpxtpx:course></gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.885058" lon="-77.114068"><ele>88.58</ele><time>2013-08-23T14:47:54Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</gpxtpx:speed><gpxtpx:course>180.71</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt lat="38.885008" lon="-77.114062"><ele>88.10</
ele><time>2013-08-23T14:47:55Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</
gpxtpx:speed><gpxtpx:course>177.88</gpxtpx:course></gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.884954" lon="-77.114056"><ele>88.10</ele><time>2013-08-23T14:47:56Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>6.86</gpxtpx:speed><gpxtpx:course>179.29</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt lat="38.884389" lon="-77.113991"><ele>85.69</
ele><time>2013-08-23T14:48:05Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>6.86</
gpxtpx:speed><gpxtpx:course>182.12</gpxtpx:course></gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.884175" lon="-77.113967"><ele>85.21</ele><time>2013-08-23T14:48:09Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</gpxtpx:speed><gpxtpx:course>184.94</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt lat="38.883740" lon="-77.114060"><ele>82.33</
ele><time>2013-08-23T14:48:19Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</
gpxtpx:speed><gpxtpx:course>254.12</gpxtpx:course></gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.883736" lon="-77.114121"><ele>82.33</ele><time>2013-08-23T14:48:20Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</gpxtpx:speed><gpxtpx:course>266.82</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt lat="38.883690" lon="-77.114860"><ele>80.41</
ele><time>2013-08-23T14:48:31Z</time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>4.12</
gpxtpx:speed><gpxtpx:course>265.41</gpxtpx:course></gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.883641" lon="-77.115661"><ele>81.37</ele><time>2013-08-23T14:48:44Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.49</gpxtpx:speed><gpxtpx:cour[Mucha ~/Desktop/Garmin]$
```

(Reformatted for improved readability)

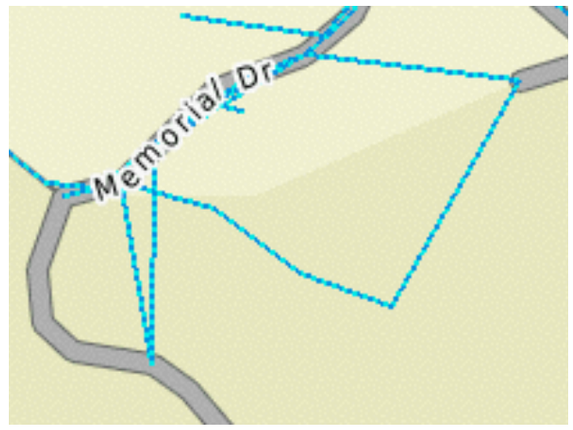
```
<trkpt lat="38.885058" lon="-77.114068">
        <ele>88.58</ele><time>2013-08-23T14:47:54Z</time>
        <gpxtpx:speed>5.49</gpxtpx:speed>
        <gpxtpx:course>180.71</gpxtpx:course>
</trkpt>

<trkpt lat="38.885008" lon="-77.114062">
        <ele>88.10</ele><time>2013-08-23T14:47:55Z</time>
        <gpxtpx:speed>5.49</gpxtpx:speed>
        <gpxtpx:course>177.88</gpxtpx:course>
</trkpt>
```

Accuracy of 0.000001º lat is ≈ 40,000 km ÷ 360 x .000001 ≈ 0.1 m

≈ 10cm

GPS accuracy is 7.8 meters w/ 95% confidence level

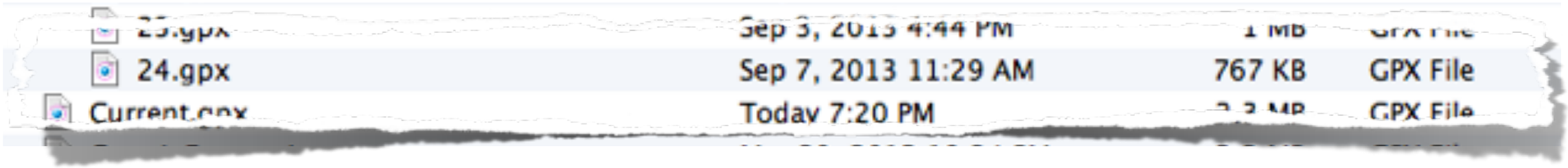
- http://www.gps.gov/systems/gps/performance/accuracy/



The Garmin GPS stores a lot of information.

Each XML record is roughly 266 bytes:

```
<ele>88.58</ele><time>2013-08-23T14:47:54Z</
time><extensions><gpxtpx:TrackPointExtension><gpxtpx:speed>5.
49</gpxtpx:speed><gpxtpx:course>180.71</gpxtpx:course></
gpxtpx:TrackPointExtension></extensions></trkpt><trkpt
lat="38.885008" lon="-77.114062"><ele>88.10</ele>
```

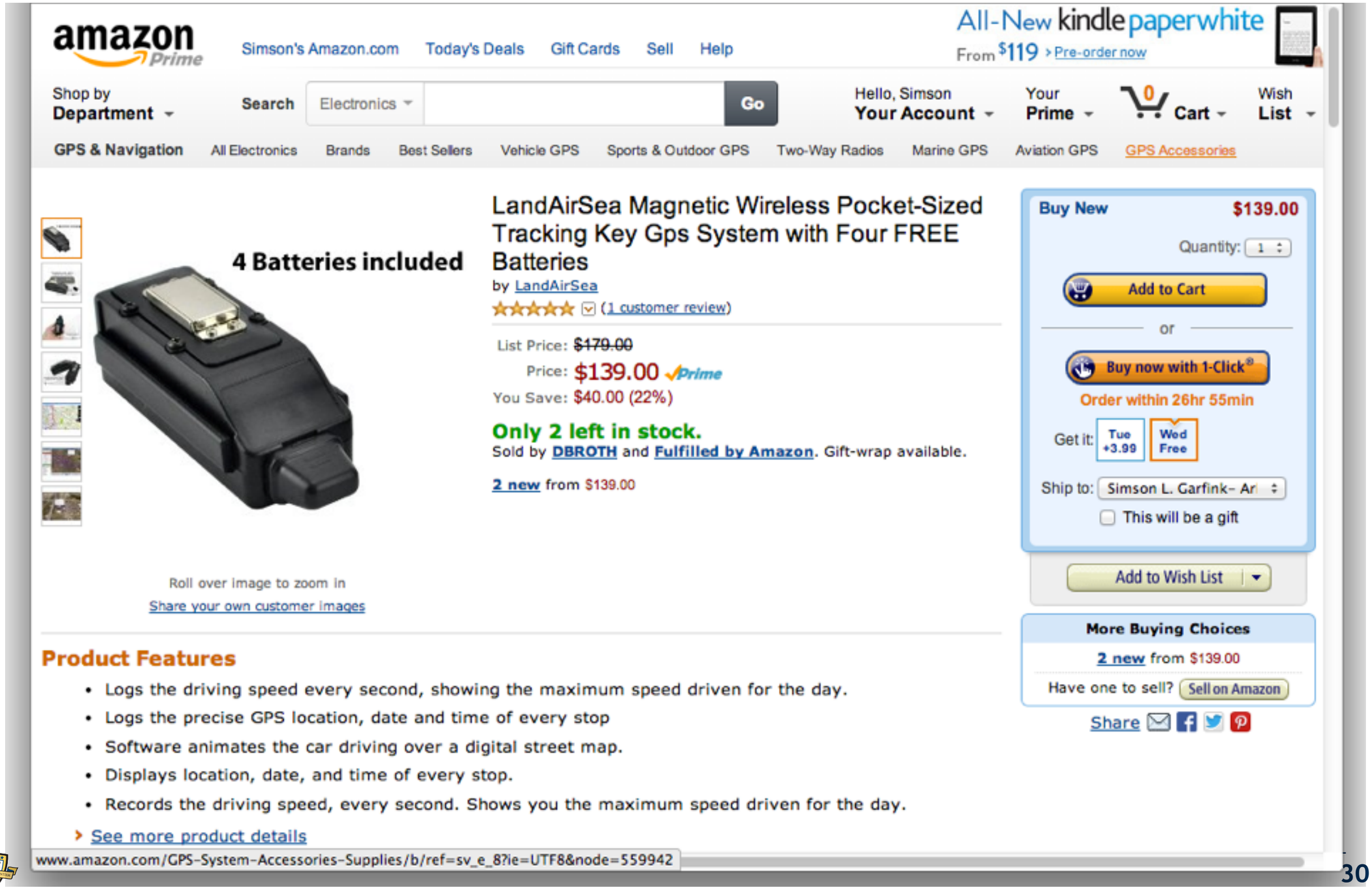


The file 24.gpx has 766,553 ÷ 266 ≈ 2,900 tracking points in it.

- My GPS has 44MB of track points from March 7, 2013 → Sept. 15, 2013
  —*That's over 150,000 tracking points!*

Inadvertently collected, but incredibly useful.

LandAirSea Magnetic Wireless Pocket-Sized Tracking Key GPS System sold by Amazon: $139.00



George Ford's wife put a LandSeaAir tracker on his car because she suspected he was having an affair.

12 WBNG ACTION NEWS

Search articles, vi

NEWS | WEATHER | SPORTS | FEATURES | LIFESTYLE | EXPERTS | VIDEO | Closi

EYES FWD

Top Stories in Local

Front Street Exit to close for two months

WATCH: Last laps at Tioga Downs for 2013 season

News » Local

Email this article | Print this article | Tweet | Like 24 | +1 0

George Ford Murder Conviction Upheld

By WBNG News

December 26, 2011

Updated Dec 26, 2011 at 2:02 PM EDT

Norwich, NY (WBNG Binghamton) The New York Supreme Court Appellate Division rejected an appeal of the murder conviction of George Ford, 46, in Chenango County.

Ford is serving a sentence of 25 years to life for 2nd degree murder.

WBNG E-News - Sign Up For Our Newsletter!

RELATED:

George Ford Appeal: Third Judical Appellate Division

Data collected by the device was instrumental in his murder conviction.

Digital forensics makes *digital evidence* available for decisions

Preparation: policy, training & tools

Collect & preserve evidence

Extract preserved data

Analysis

Reporting & Testimony

Digital Forensics “research” traditionally focused on training, collection & extraction

Preparation: policy, training & tools

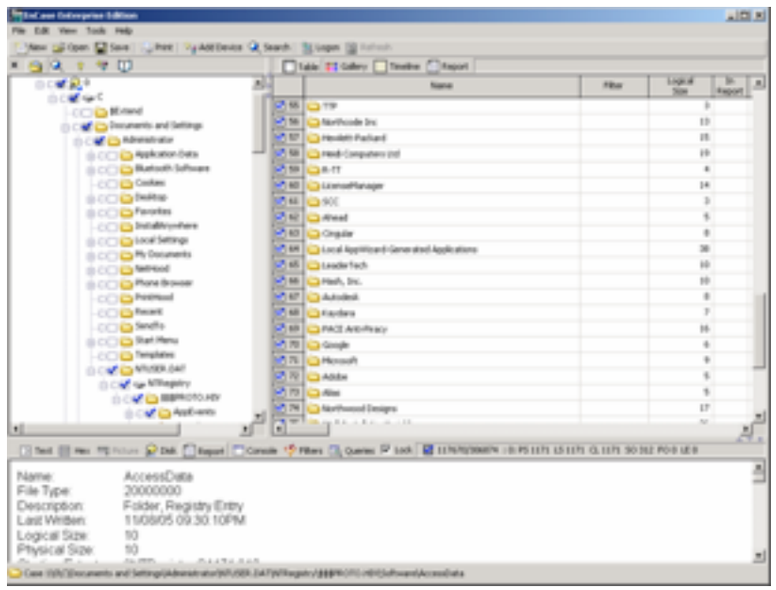
Collect & preserve evidence

Extract preserved data

Training the force

CyberCorps®
Defending America's Cyberspace

Collection Techniques



Reverse engineering

Most tools were developed for law enforcement.

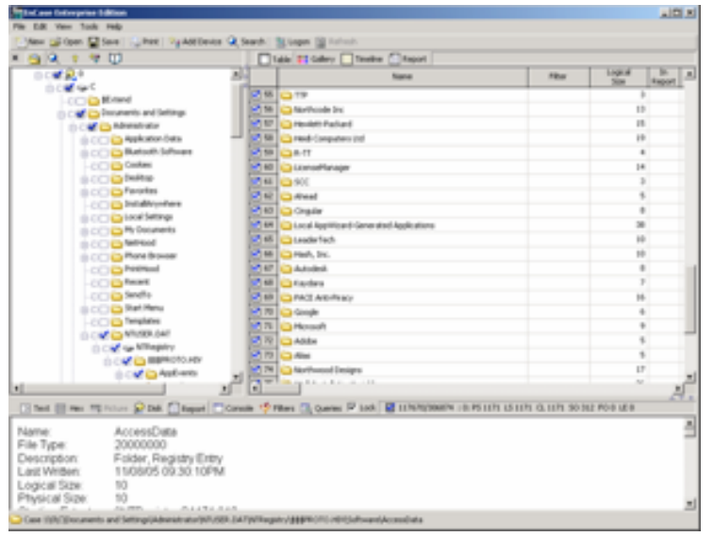
Tools follow a simple model for manual analysis:

- Extract data
- Make data visibility
- Simple string search

Hard Drive from desktop

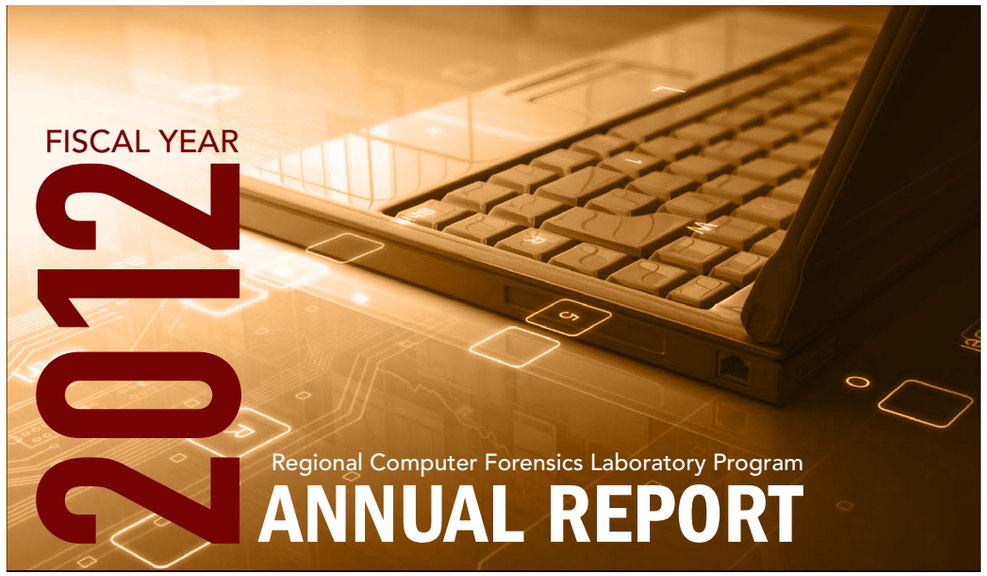
“Write Blocker”

Extraction tools



Encase Forensic

Collected data increases dramatically each year.



FY 12 ACCOMPLISHMENTS

BY THE NUMBERS

	FY 12	FY11	% CHANGE
DIGITAL FORENSIC SERVICES			
Agency Requests No. of agencies that received RCFL assistance	842	766	**10%**
Service Requests No. of requests for assistance received by RCFLs	5,060	6,318	**-20%**
Examinations Completed No. of examinations completed at RCFLs	8,566	7,629	**12%**
Terabytes Processed A terabyte (TB) is a unit of measurement for data storage capacity equivalent to 1,024 gigabytes. One TB is roughly equivalent to the information in 1,000 encyclopedias.	5,986	4,263	**40%**
Field Services No. of onsite operations conducted by law enforcement for which RCFLs provided assistance	553	689	**-20%**
Examiner Testimony in Court No. of times RCFL Examiners testified in court and/or at hearings. This does not include subpoenas to testify when testimony was not required.	101	97	**4%**
KIOSK SERVICES			
Total Kiosk Use No. of times law enforcement officers used the CPIK, LMK, and VCPK	13,556	8,553	**58%**
Cell Phone Investigative Kiosk (CPIK) Use No. of times law enforcement officers used the CPIK to review cellular phones at RCFLs in support of investigations	8,795	5,956	**48%**
Loose Media Kiosk (LMK) Use No. of times law enforcement officers used the LMK to review loose media at RCFLs in support of investigations	3,665	1,683	**118%**
Virtual Cell Phone Kiosk (VCPK) Use No. of times law enforcement officers used the VCPK (only available through the HARCFL) to remotely review cellular telephones from their agencies in support of investigations)	1,096	914	**20%**

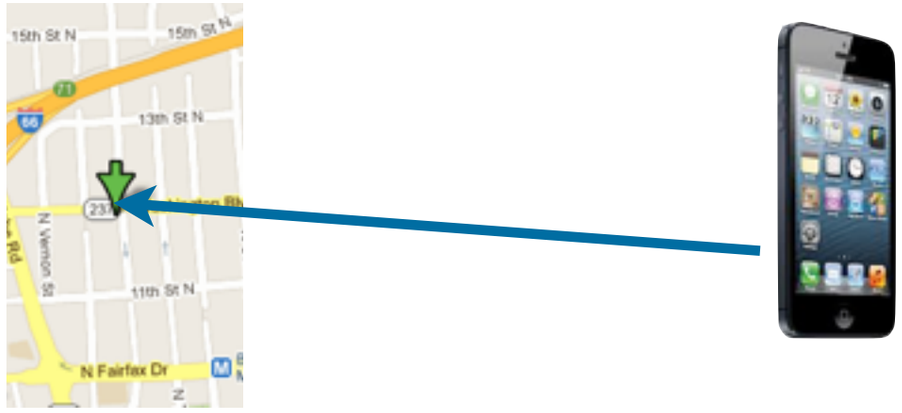
My work focuses on developing better approaches for automation and analysis — “big data for little devices”

Automatically identify high-value data.

—*Contacts, calendar, documents*

—*Software*

—*GPS / Temporal*



Correlate — devices with identical or similar copies.

- previously unknown organizations or networks
- data/threats that are unusual or emerging

Presentation and Integration:

- Make the results understandable.
- Effect organizational change through adoption & integration

Human Language Technology

- Apply to العربية ,עִבְרִית, español, 汉语/漢語, 日本語, svenska, etc.

Three principles underly this research.

1. Work with “big data.”
   - “Big data” is our advantage — use it!
   - Many techniques developed on small systems don’t scale

Three principles underly this research.

1. Work with “big data.”
2. Automation is essential.
   - Today most forensic analysis is done manually.
   - We are developing techniques & tools to allow automation.

Three principles underly this research.

1. Work with “big data.”
2. Automation is essential.
3. Concentrate on *bulk data*.
   - Leverage data that are fragmented and incomplete
     - —*Deleted and partially overwritten files*
     - —*Fragments of memory in swap & hibernation*
     - —*Tool marks*

—*Sencar & Memon,*”Identification and recovery of JPEG files with missing fragments,” DFRWS 2009

Example: Integrating Human Language Technology with Digital Forensics

Problems:

- Forensic examiners spend significant time looking for “unusual” file names.
- Many of the file names are not in English.

Solution:

- Model of what makes file names “unusual.”
- Translate non-English file names into English.

New problem: path names are ambiguous and frequently multi-lingual.

- Documents and Settings/defaultuser/ Mes documents/Ma musique/Desktop.ini
- Mis Documentos/SalvadorJP/Excel/ GRUPOS.xls
- Documents and Settings/3742008/ Configuración local/Datos de programa/ Microsoft/Internet Explorer/.
- top.com/السلسلة المعلوماتية/تصميماتي.jpg
  becomes:
  top.com/My designs/The computer-based series.jpg

—*Rowe, Neil, Schwamm, Riqui, Garfinkel, Simson. Language Translation for File Paths, DFRWS 2013, Aug 4-7, 2013. Monterey, CA. BEST PAPER AWARD*

We do science with “real data.”

The Real Data Corpus (60TB)

- Disks, camera cards, & cell phones purchased on the secondary market.
- Most contain data from previous users.
- Mostly acquire outside the US:
  —*Canada, China, England, Germany, France, India, Israel, Japan, Pakistan, Palestine, etc.*
- Thousands of devices (HDs, CDs, DVDs, flash, etc.)

Mobile Phone Application Corpus

- Android Applications; Mobile Malware; etc.

The problems we encounter obtaining, curating and exploiting this data mirror those of national organizations

—*Garfinkel, Farrell, Roussev and Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, DFRWS 2009. BEST PAPER AWARD.*

—*http://digitalcorpora.org/*

We manufacture data that can be freely redistributed.

Files from US Government Web Servers (500GB)

- ≈1 million heterogeneous files
  - —*Documents (Word, Excel, PDF, etc.); Images (JPEG, PNG, etc.)*
  - —*Database Files; HTML files; Log files; XML*
- Freely redistributable; Many different file types
- This database was surprising difficulty to collect, curate, and distribute:
  - —*Scale created data collection and management problems.*
  - —*Copyright, Privacy & Provenance issues.*

Advantage over flickr & youtube: persistence & copyright

<abstract>NOAA's National Geophysical Data Center (NGDC) is building high-resolution digital elevation models (DEMs) for select U.S. coastal regions. … </abstract>

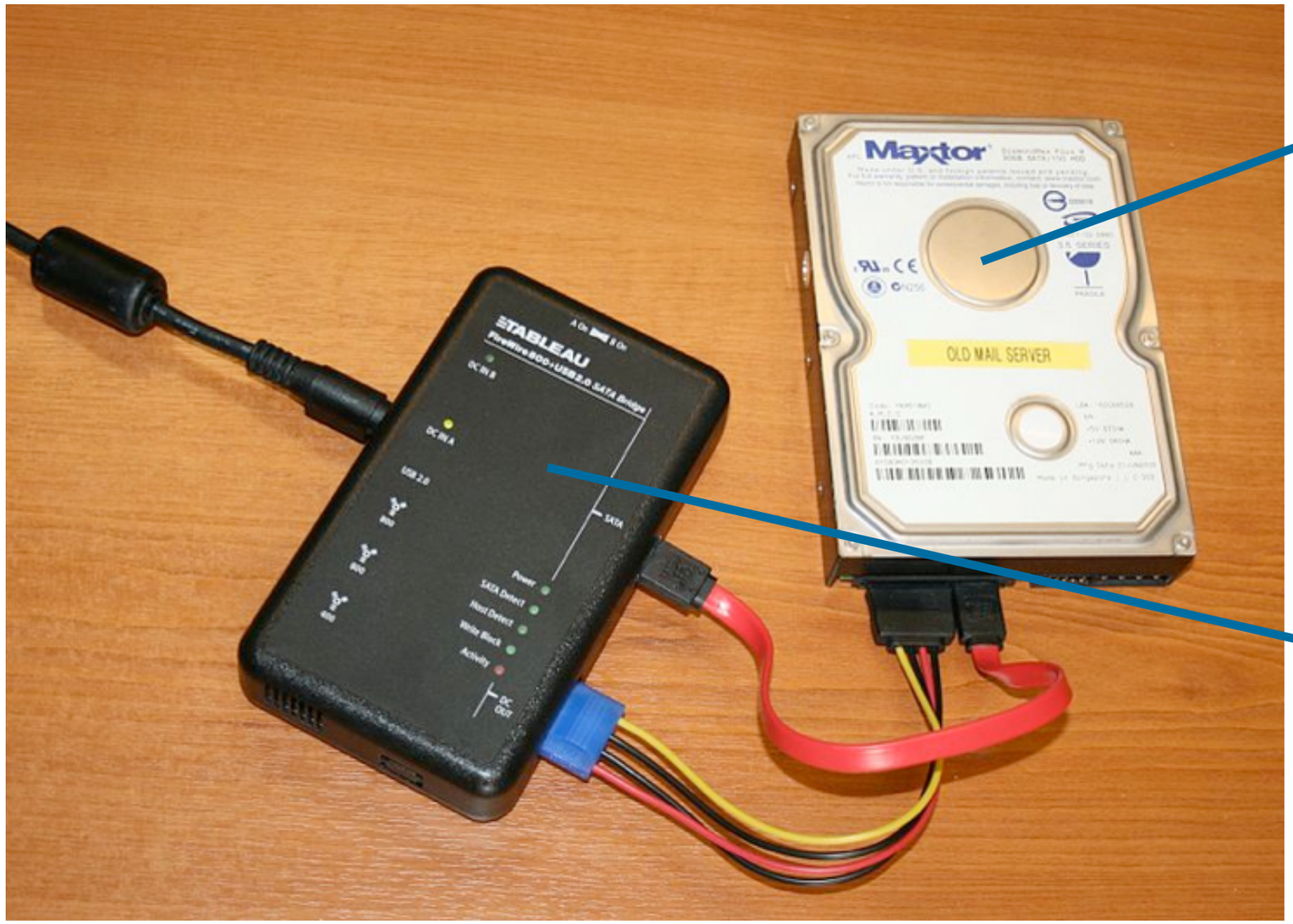
<abstract>This data set contains data for birds caught with mistnets and with other means for sampling Avian Influenza (AI)….</abstract>

—*http://digitalcorpora.org/*

Challenges Facing Digital Forensics

Extracting digital evidence was simple five years ago.

"Imaging tools" extracted data without modification.



"Write Blocker" prevents accidental overwriting.

Original device stored in evidence locker.

Forensic copy ("disk image") stored on a storage array.

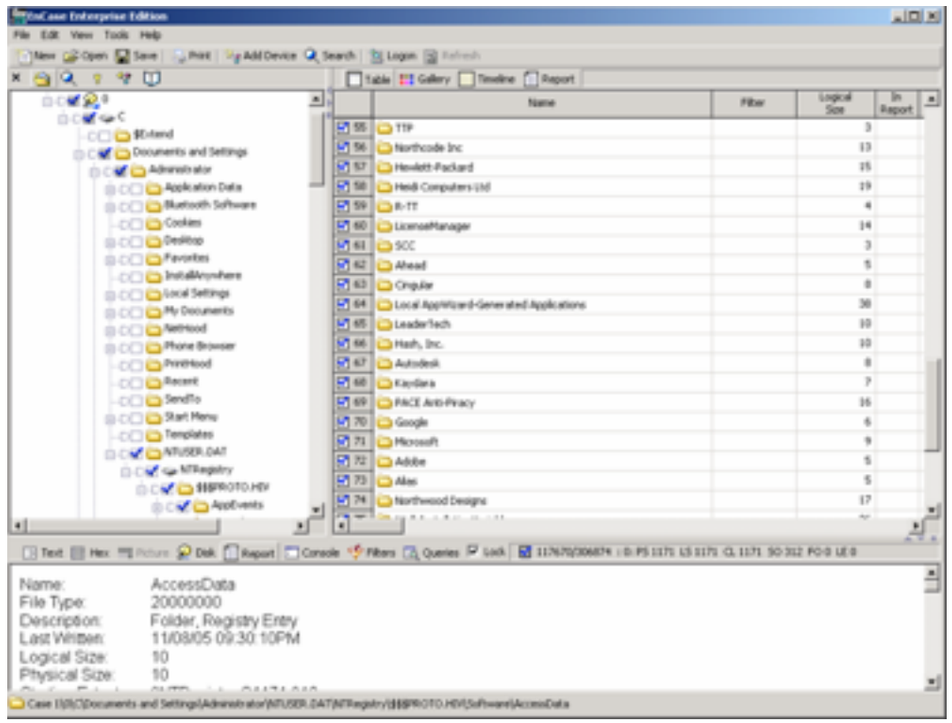
Analyzing digital evidence was simple five years ago.

Commercial tools extracted *files* from disk images

- Display of *allocated* & *deleted* files.
- String search
- File extraction
- File “carving”
- Examining disk sectors

Job of analyst:

- Find interesting data
- Report on it.



Today Digital Forensics is faced with 5 crippling challenges:

1. Device Diversity





2. Data Diversity


3. Data Scale

SanDisk Ultra 64 GB MicroSDXC Class 10 UHS-1 Memory Card with Adapter (SDSDQU-064G-AFFP-A)

~~$99.99~~ **$55.33**

Seagate Desktop HDD 4 TB SATA 6Gb/s NCQ 64MB Cache 3.5-Inch Internal Bare Drive ST4000DM000

~~$209.99~~ **$189.99**

4. Human Capital

→

5. Cloud & Encryption

Mobile devices exhibit multiple challenges.

Operating system:

- Android? iPhone? Blackberry? Feature Phone?

Access to the data:

- PIN lock?
- Encrypted Storage?
- Stored locally or in the cloud?

Applications:

- Built-in? Downloaded from “App Store”?
- Custom-written?
- Self-destruct / remote wipe?
- Malware?

Human Language: English? Korean? Chinese?

This $12 phone from Hong Kong could contain important evidence.

Spec	Phone
Price	$12
CPU	260 MHz, 32-bit
RAM	8MiB
Interfaces	USB, microSD, SIM
Wireless	Quadband GSM, Bluetooth
Power	Li-Poly battery, includes adapter
Display	Two-color OLED

http://www.bunniestudios.com/blog/?p=3040

The “CSI Effect” creates unrealistic expectations.

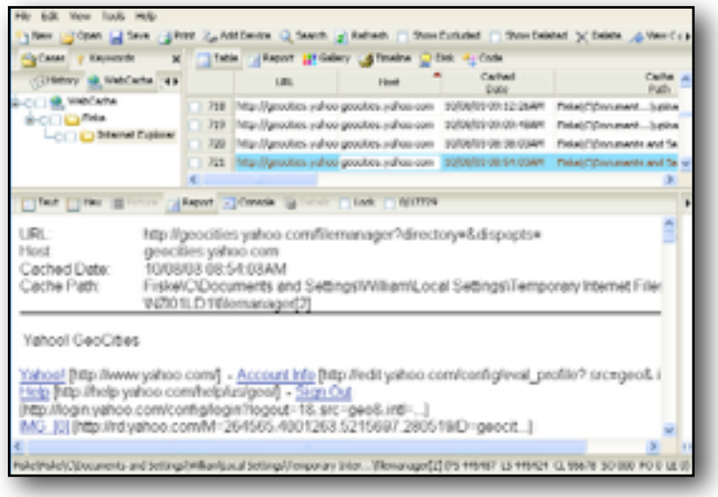
TV digital forensics:

- Every investigator is trained on every tool
- Correlation is easy and instantaneous
- There are no false positives
- Overwritten data can be recovered
- Encrypted data can usually be cracked
- It is impossible to delete anything

The reality:

- Overwritten data cannot be recovered
- Encrypted data usually can't be decrypted
- Forensics rarely answers questions or establishes guilt
- Tools crash a lot

—Digital Forensics: a difficult process that looks easy



EnCase

Digital Forensics must respond with new science.

Current approaches don't scale

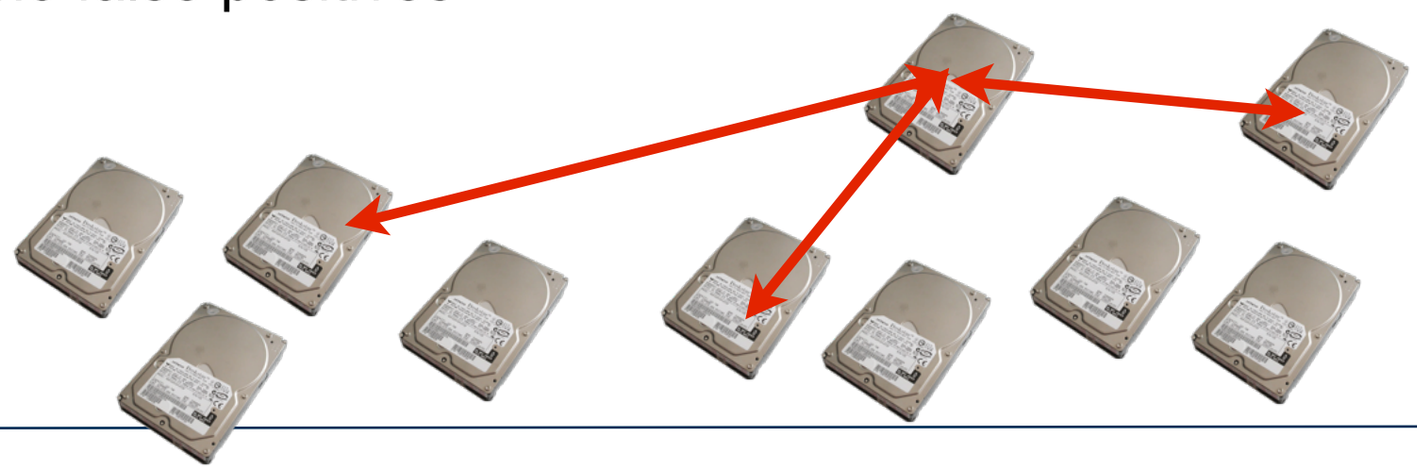
- User spent *years* assembling email, documents, etc.
- Analysts have days or hours to process it
- Police analyze top-of-the-line systems

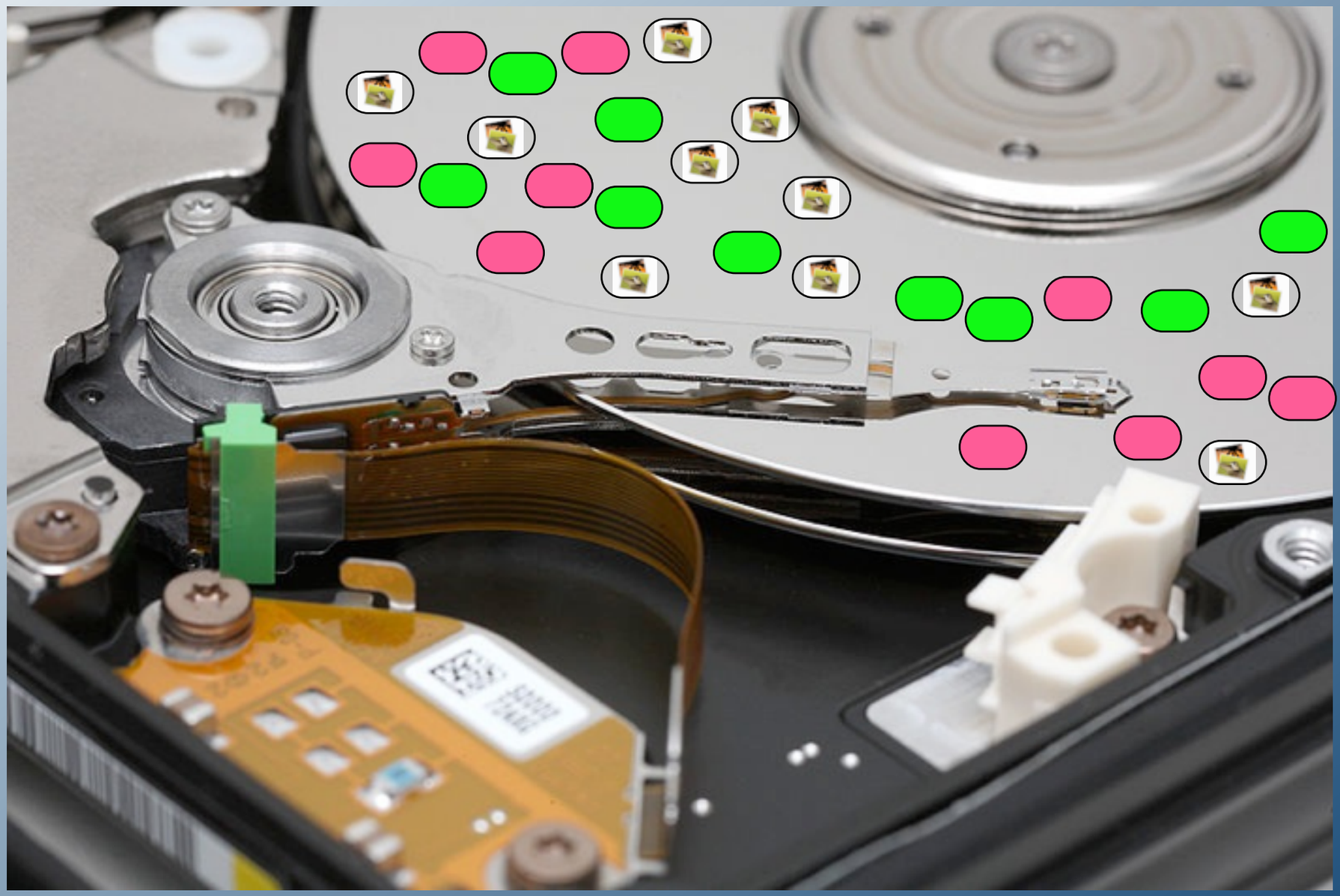
  —*with top-of-the-line systems*
- National Labs have large-scale server farms

  —*to analyze huge collections*

Our approach: leverage our massive data advantage

- Outlier detection and correlation
- Operate autonomously on incomplete, heterogeneous datasets
- Automatically recalibrate; no false positives





The Triage Problem

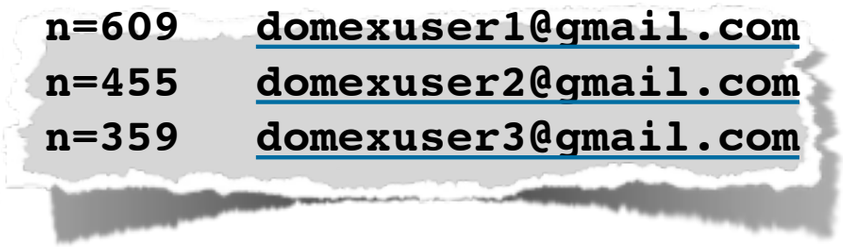
Digital media triage:
Deciding where to start

Imagine you encounter a large number of computers, USB drives, etc.

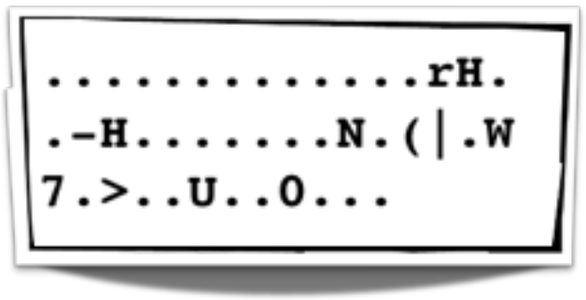
Where do you start?

We have developed many triage techniques.

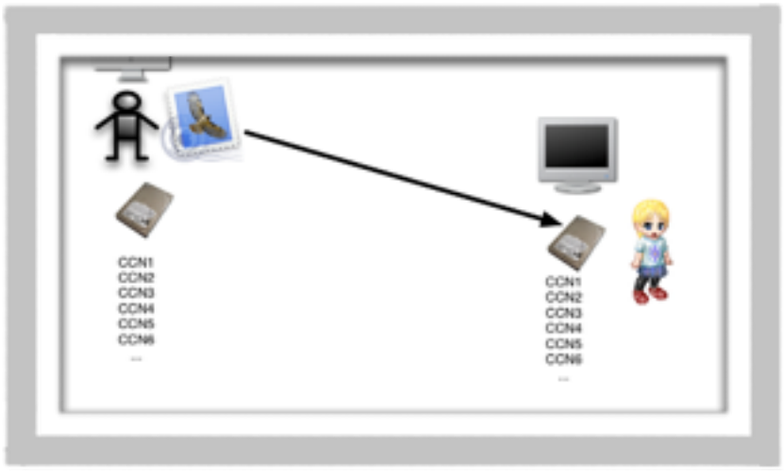
1. Histogram analysis
   —*We count the results*



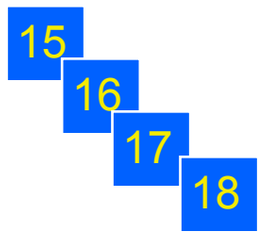
2. Optimistic decompression
   —*We try to decompress the data*



3. Cross-drive analysis
   —*We look for two computers with the same email address*



4. sector hashing
   —*We look at individual disk sectors.*



5. Random sampling
   —*We look at 1% of the data, randomly chosen*

Email addresses are powerful digital forensic identifiers

Email addresses can reveal:

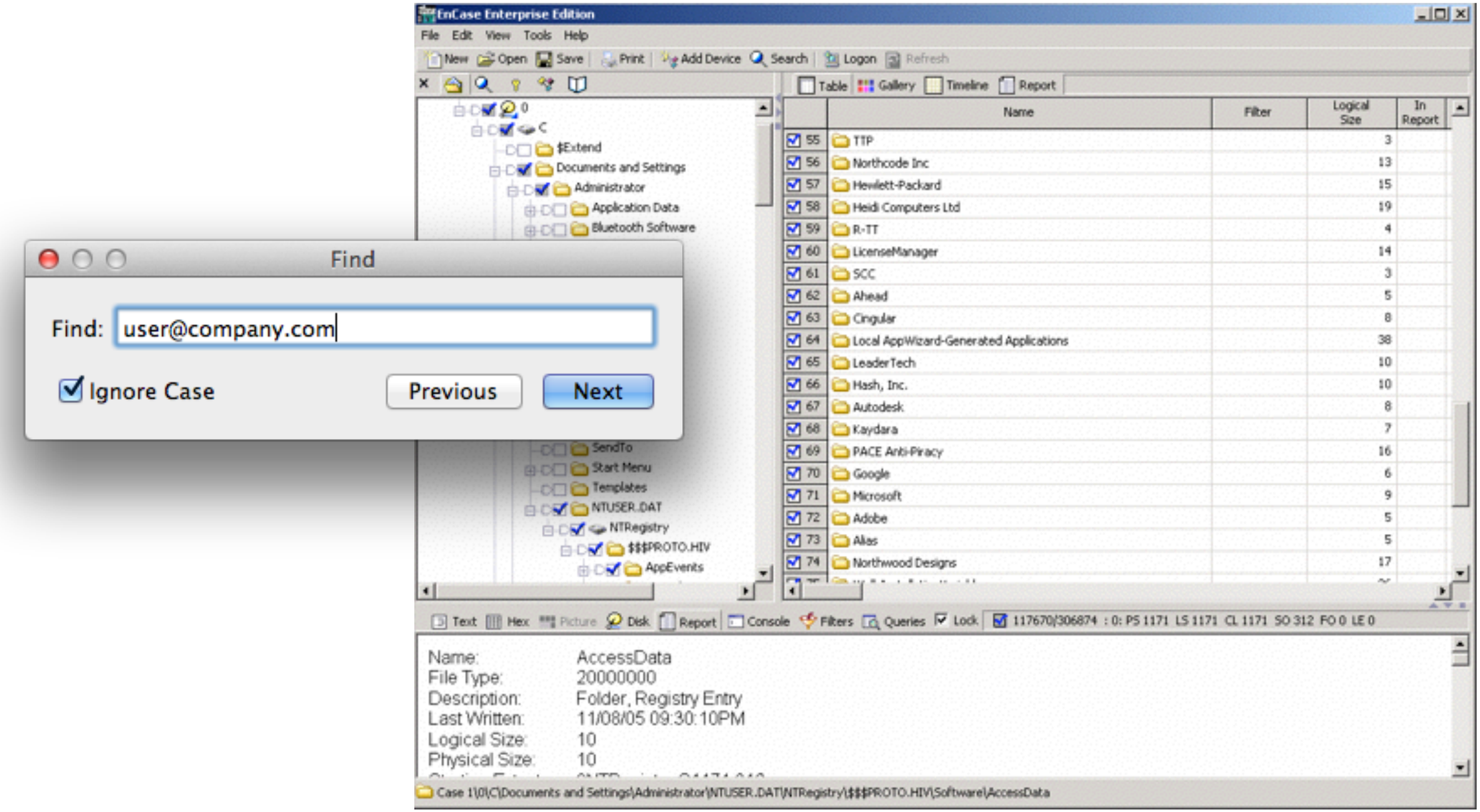
- User(s) of a device
- Associates
- Connections between devices

ABC@company.com
DEF@company.com
XYZ@company.com

HIJ@network.net
KLM@network.net
NOP@network.net
XYZ@company.com

- Today's forensic tools implement two strategies for extracting email addresses.
  1. *Text extraction from files*
  2. *Text extraction from bulk data*

But most digital forensics tools find email addresses with string search



With string search, you only find what you are looking for.

bulk_extractor extracts email addresses.

For each email address, reports location, email address, & context

```
34427974 grafta@bl.com          \x028\x00\x08\x01.\x01\xE4c\x00
24900678 grafta@bl.com          \x028\x00\x08\x01.\x01\xE4c\x00
50392739 inet@microsoft.com     ica Server" by "
26735686 grafta@bl.com          \x028\x00\x08\x01.\x01\xE4c\x00
39265456 domexuser2@gmail.com   tocol>\x0D\x0A\x09\x09<n
39267100 domexuser2@live.com    tocol>\x0D\x0A\x09\x09<na
39269992 domexuser1@gmail.com   il - - domexuser1@gmail.com
39270105 domexuser1@gmail.com   l - Inbox (1) - domexus
40893040 domexuser2@live.com    tocol>\x0D\x0A\x09\x09<name>
40948912 domexuser2@gmail.com   tocol>\x0D\x0A\x09\x09<name>
40950441 domexuser2@live.com    tocol>\x0D\x0A\x09\x09<name>
51781228 dbaron@dbaron.org      L. David Baron <dbaron@dbaron.org
51788157 bzbarsky@mit.edu       oris Zbarsky\x0A# <bzbarsky@mit.e
51789901 bzbarsky@mit.edu       oris Zbarsky\x0A# <bzbarsky@mit.e
```

Histogram analysis shows the important email addresses.

```
n=609    domexuser1@gmail.com   (utf16=303)
n=455    domexuser2@gmail.com   (utf16=225)
n=359    domexuser3@gmail.com   (utf16=204)
n=268    ips@mail.ips.es
n=252    premium-server@thawte.com
n=243    cps-requests@verisign.com     (utf16=3)
n=243    someone@example.com    (utf16=234)
n=221    domexuser2@live.com    (utf16=59)
n=198    domexuser1@hotmail.com(utf16=80)
n=185    domexuser1@live.com    (utf16=59)
n=175    domexuser2@hotmail.com(utf16=97)
n=145    inet@microsoft.com
n=115    example@passport.com   (utf16=115)
n=115    myname@msn.com          (utf16=115)
n=94     info@valicert.com
n=91     piracy@microsoft.com   (utf16=91)
n=80     certificate@trustcenter.de
n=78     name_123@hotmail.com   (utf16=78)
n=74     talkback@mozilla.org   (utf16=12)
n=69     hewitt@netscape.com    (utf16=1)
n=64     lord@netscape.com
```

Primary Users

From installed software

Notice that we combine UTF8 and UTF16

It's important to distinguish email addresses that are relevant to a case from those that are not.

The #4 address is ips@mail.ips.es

- We should probably ignore these

```
n=609   domexuser1@gmail.com
n=455   domexuser2@gmail.com
n=359   domexuser3@gmail.com
n=268   ips@mail.ips.es
n=252   premium-server@thawte.com
n=243   cps-requests@verisign.com
n=243   someone@example.com
n=221   domexuser2@live.com
n=198   domexuser1@hotmail.com
n=185   domexuser1@live.com
```

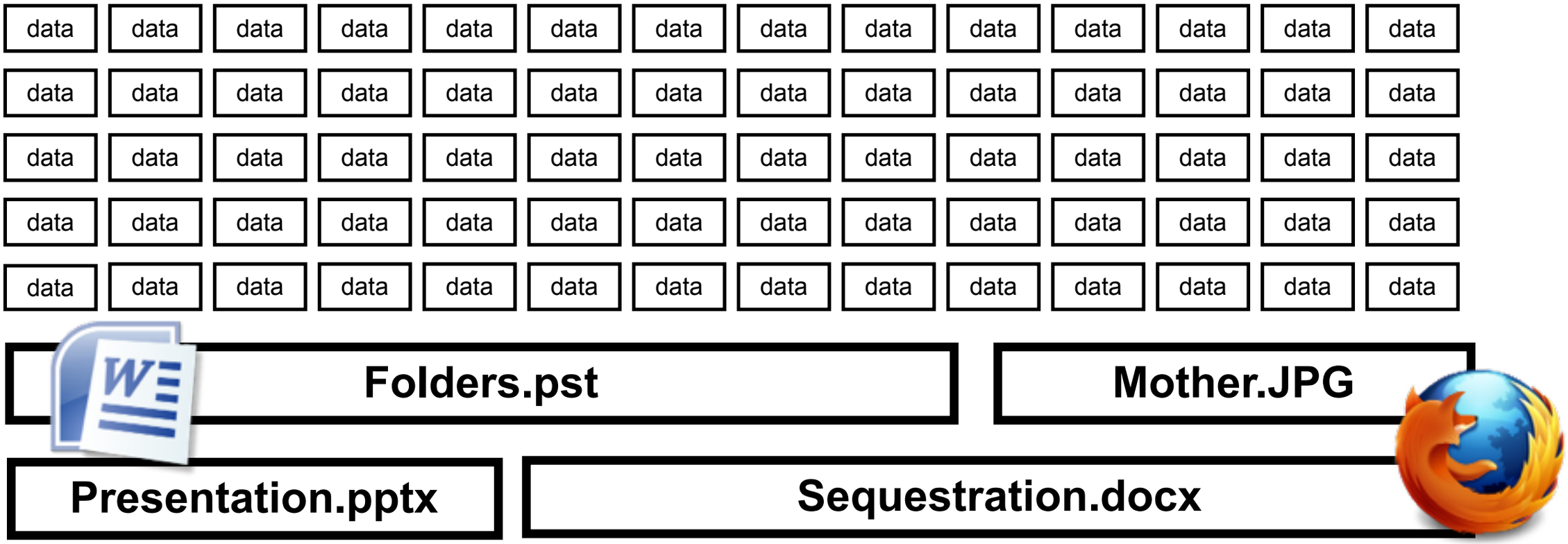
Other sources that to ignore:

- Windows binaries; SSL certificates; Sample documents; News Stories

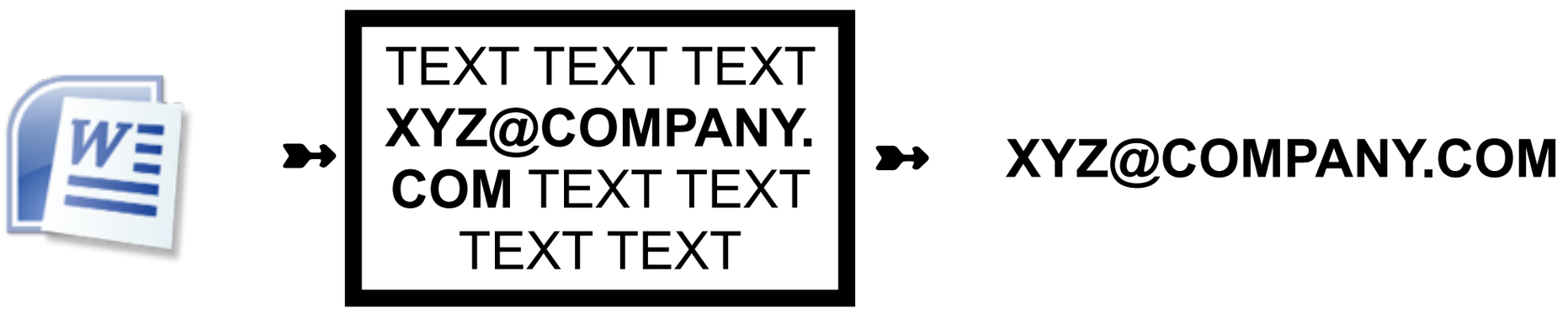
Ignore lists are expensive to maintain.

- Find them automatically by analyzing multiple drives!
- Email addresses on thousands of different drives are probably irrelevant.

Storage devices arrange files in individual sectors.



Email addresses might be in any file, any sector:



Every email address is a sequence of bytes.

A simple email address:

`XYZ@company.com`

Stored on disk / in memory as 15 bytes:

`X Y Z @ c o m p a n y . c o m`

Each byte is 8-bits. Range is 0-255

`88 89 90 64 99 111 109 112 97 110 121 46 99 111 109`

Normally bytes are displayed in hexadecimal notation:

`58 59 5a 40 63 6f 6d 70 61 6e 79 2e 63 6f 6d`

This is called UNICODE (UTF-8)

Every email address is a sequence of bytes.

A simple email address:

XYZ@company.com

Stored on disk / in memory as 15 bytes:

X	Y	Z	@	c	o	m	p	a	n	y	.	c	o	m

Each byte is 8-bits. Range is 0-255

88	89	90	64	99	111	109	112	97	110	121	46	99	111	109

Normally bytes are displayed in hexadecimal notation:

58	59	5a	40	63	6f	6d	70	61	6e	79	2e	63	6f	6d

This is called UNICODE (UTF-8)

Some email addresses are easy to spot

data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data

Folders.pst

Mother.JPG

Presentation.pptx

Sequestration.docx

```
a097 83a1 ed96 26a6 3c69 3d0f 750a 2399   ......&.<i=.u.#.
a2b5 bea7 692f 5847 a38a dd53 082c add5   ....i/XG...S.,..
5061 b64c 721d 864b 90b6 b55f bb04 735c   Pa.Lr..K..._..s\
9448 6730 5453 df64 813e b603 5795 2242   .Hg0TS.d.>..W."B
e9c8 7454 7322 7cdc b60e 97af 2f64 2728   ..tTs"|...../d'(
3c58 595a 4043 4f4d 5041 4e59 2e43 4f4d   <XYZ@COMPANY.COM
a9e9 e92c a3f8 6e46 0530 8a88 c7a2 5d2b   ...,..nF.0....]+
d89d 77cc fe1e f637 f3f3 d0af 1b47 c09b   ..w....7.....G..
```

Email addresses like these can be found with **regular expressions**

data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data
data	data	data	data	data	data	data	data	data	data	data	data	data	data

Folders.pst

Mother.JPG

Presentation.pptx

Sequestration.docx

```
a097 83a1 ed96 26a6 3c69 3d0f 750a 2399   ......&.<i=.u.#.
a2b5 bea7 692f 5847 a38a dd53 082c add5   ....i/XG...S.,..
5061 b64c 721d 864b 90b6 b55f bb04 [illegible]35c   Pa.Lr..K..._..s\
9448 6730 5453 df64 [illegible]42   .Hg0TS.d.>..W."B
e9c8 7454 7322 7cdc b[illegible]   ..tTs"|...../d'(
3c58 595a 4043 4f4d 5[illegible]   <XYZ@COMPANY.COM
a9e9 e92c a3f8 6e46 0[illegible]   ...,..nF.0....]+
d89d 77cc fe1e f637 f3f3 d0af 1b47 [illegible]9b   ..w....7.....G..
```

XYZ@company.com

Problem: byte sequences can be encoded in many ways.

XYZ@company.com

- Unicode: “XYZ@company.com”

```
58 59 5a 40 63 6f 6d 70 61 6e 79 2e 63 6f 6d
```

- Base 16: “58595a40636f6d70616e792e636f6d0a”

```
3538 3539 3561 3430 3633 3666 3664 3730  58595a40636f6d70
3631 3665 3739 3265 3633 3666 3664 3061  616e792e636f6d0a
```

- Base 64: “WFlaQGNvbXBhbnkuY29tCg===”

```
5746 6c61 5147 4e76 6258 4268 626e 6b75  WFlaQGNvbXBhbnku
5932 3974 4367 3d3d 3d0a                 Y29tCg===.
```

- Compression: echo “XYZ@company.com” | compress | xxd

```
1f9d 9058 b268 0132 e64d 1b38 61dc e471  ...X.h.2.M.8a..q
51b0 8d02                                Q...
```

Compression works by eliminating repeated sequences:

Computers use compression to save memory:

```
5859 5a40 636f 6d70 616e 792e 636f 6d20  XYZ@company.com
4142 4340 636f 6d70 616e 792e 636f 6d20  ABC@company.com
4445 4640 636f 6d70 616e 792e 636f 6d20  DEF@company.com
```

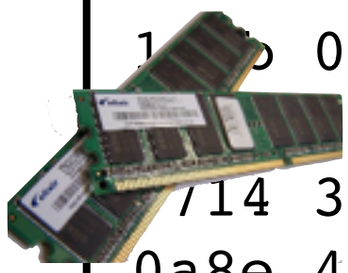
Compressed with “gzip:”

```
1f8b 0800 0000 0000 0203 8b88 8c72 48ce  .............rH.
cf2d 48cc abd4 03d2 0a8e 4ece 287c 1757  .-H.......N.(|.W
3714 3e00 b455 c1c5 3000 0000            7.>..U..0...
```

Compressed email addresses do not “look” like email addresses!

—*You can’t find them with regular expressions!*

It’s hard to see compressed email address in bulk data.



```
e327 962d 6450 3d91 c945 3bed 97a6 a4cd   .'.-dP=..E;.....
1[illegible]b 0800 0000 0000 0203 8b88 8c72 48ce   .............rH.
[illegible]8cc abd4 03d2 0a8e 4ece 287c 1757   .-H.......N.(|.W
[illegible]714 3e00 b455 c1c5 3000 0000 0000 0000   7.>..U..0.......
0a8e 4ece 287c 1757 3714 3e00 a175 10ed   ..N.(|.W7.>..u..
```

Folders.pst

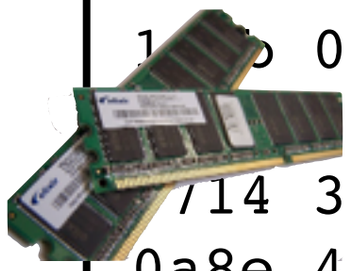
Mother.JPG

Presentation.pptx

Sequestration.docx

```
a097 83a1 ed96 26a6 3c69 3d0f 750a 2399   ......&.<i=.u.#.
a2b5 bea7 692f 5847 a38a dd53 082c add5   ....i/XG...S.,..
5061 b64c 721d 864b 90b6 b55f bb04 735c   Pa.Lr..K..._..s\
9448 6730 5453 df64 813e b603 5795 2242   .Hg0TS.d.>..W."B
[illegible]8 7454 7322 7cdc b60e 97af 2f64 2728   ..tTs"|...../d'(
[illegible]4bd 2a84 2dfe 50ea 5935 c349 1513   <XYZ@COMPANY.COM
[illegible]e92c a3f8 6e46 0530 8a88 c7a2 5d2b   ...,..nF.0....]+
d89d 77cc fe1e f637 f3f3 d0af 1b47 c09b   ..w....7.....G..
```

It's hard to see compressed email address in bulk data.



```
e327 962d 6450 3d91 c945 3bed 97a6 ..cd   .'.-dP=..E;.....
..b0 0800 0000 0000 0...                  ..............rH.
....  8cc abd4 03d2 0...                  .-H.......N.(|.W
..14 3e00 b455 c1c5 3...                  7.>..U..0.......
0a8e 4ece 287c 1757 3714 3e00 a175 ..ed   ..N.(|.W7.>..u..
```

XYZ@company.com
ABC@company.com
DEF@company.com

Folders.pst

Mother.JPG

Presentation.pptx

Sequestration.docx

```
a097 83a1 ed96 26a6 3c69 3d0f 750a 2399   ......&.<i=.u.#.
a2b5 bea7 692f 5847 a38a dd53 082c add5   ....i/XG...S.,..
5061 b64c 721d 864b 90b6 b55f bb04 735c   Pa.Lr..K..._..s\
9448 6730 5453 df64 813e b603 5795 2242   .Hg0TS.d.>..W."B
e..8 7454 7322 7cdc b60e 97af 2f64 2728   ..tTs"|...../d'(
....  4bd 2a84 2dfe 50ea 5935 c349 1513   <XYZ@COMPANY.COM
....  e92c a3f8 6e46 0530 8a88 c7a2 5d2b   ...,..nF.0....]+
d89d 77cc fe1e f637 f3f3 d0af 1b47 c09b   ..w....7.....G..
```

bulk_extractor breaks the disk into individual blocks.

e327 962d 6450 3d91	c945 3bed 97a6 a4cd	.'.-dP=..E;.....
1f8b 0800 0000 0000	0203 8b88 8c72 48ce	.............rH.
cf2d 48cc abd4 03d2	0a8e 4ece 287c 1757	.-H.......N.(|.W
3714 3e00 b455 c1c5	3000 0000 0000 0000	7.>..U..0.......
0a8e 4ece 287c 1757	3714 3e00 a175 10ed	..N.(|.W7.>..u..

Folders.pst | **Mother.JPG**

Presentation.pptx | **Sequestration.docx**

a097 83a1 ed96 26a6	3c69 3d0f 750a 2399	......&.<i=.u.#.
a2b5 bea7 692f 5847	a38a dd53 082c add5	....i/XG...S.,..
5061 b64c 721d 864b	90b6 b55f bb04 735c	Pa.Lr..K..._..s\
9448 6730 5453 df64	813e b603 5795 2242	.Hg0TS.d.>..W."B
e9c8 7454 7322 7cdc	b60e 97af 2f64 2728	..tTs"|...../d'(
3cfb 84bd 2a84 2dfe	50ea 5935 c349 1513	<XYZ@COMPANY.COM
a9e9 e92c a3f8 6e46	0530 8a88 c7a2 5d2b	...,..nF.0....]+
d89d 77cc fe1e f637	f3f3 d0af 1b47 c09b	..w....7.....G..

Each block is tested for compressed data.
Within each block, we check each byte.

e327 962d 6450 3d91	c945 3bed 97a6 a4cd	.'.-dP=..E;.....
1f8b 0800 0000 0000	0303 8b80 8e72 48ce	.....rH.
cfad 48cc ab04 03d2	0a8e 4ece 287c 1757	..H.....N.(|.W
3714 3e00 b455 c1c5	3000 0000 0000 0000	7.>..U..0.....
0a8e 4ece 287c 1757	3714 3e00 a175 10ed	..N.(|.W7.>..u..

TEST 1 TEST 2 TEST 3

Folders.pst Mother.JPG

a097 83a1 ed96 26a6	3c69 3d0f 750a 2399	......&.<i=.u.#.
a2b5 bea7 692f 5847	a38a dd53 082c add5	....i/XG...S.,..
5001 5048 721d 864b	90b6 e551 bb04 7ac	P.PHr....K...Q..z
9448 6730 5453 df64	813e b603 5795 2242	.Hg0TS.d.>..W."B
e9c8 7454 7322 7cdc	b60e 97af 2f64 2728	..tTs"|...../d'(
3cfb 84bd 2a84 2dfe	50ea 5935 c349 1513	<XYZ@COMPANY.COM
a9e9 e52c a3f8 6e46	0500 8886 c7a2 5b0b	...,..nF......[.
d89d 77cc fe1e f637	f313 d0af 1b47 c09b	..w....7.....G..

TEST 4 TEST 5 TEST 6

TEST 7 TEST 8 TEST 9

TEST A TEST B TEST C

bulk_extractor run on a disk from India (IN10-0138) found many compressed email addresses.

Encoding	count	1) Plain in Files	2) Comp. in Files	3) Plain in non-	4) Comp in non-
Cleartext		358	--	5341	--
All Comp		--	9	--	135
GZIP	50	13	1	22	14
HIBER	39	6	1	27	5
HIBER-GZIP	23			21	2
PDF	88	1		9	78
ZIP	28	2	5	3	18
ZIP-PDF	18				18

135 out of 5700 email addresses are invisible to other forensic tools.

We extended the analysis to 1,646 disk images and many codings.

Coding	Drives	Emails	max
1) Plain in files	739	81,920	4,206
2) Comp in files	355	19,711	5,454
3) Plain in non-files	860	1,956,059	178,073
4) Comp in non-files	474	165,481	59,376
BASE64 Comp	54	219	50
BASE64-GZIP Comp	2	64	37
GZIP Comp	234	66,195	9,103
GZIP-BASE64 Comp	7	44	11
GZIP-GZIP Comp	15	12,663	11,845
GZIP-GZIP-BASE64 Comp	2	38	30
GZIP-GZIP-GZIP Comp	4	58	38
GZIP-GZIP-ZIP Comp	1	12	12
GZIP-PDF Comp	5	38	30
GZIP-ZIP Comp	6	49	30
HIBER Comp	79	1,433	217
PDF Comp	162	2,352	238
ZIP Comp	388	85,252	59,369
ZIP-BASE64 Comp	5	30	13
ZIP-BASE64-GZIP Comp	2	65	38
ZIP-GZIP Comp	14	261	132
ZIP-PDF Comp	26	115	18

We can scan for credit card numbers (CCNs), phone numbers, addresses, and other structured info.

We use multiple filters to minimize false positives

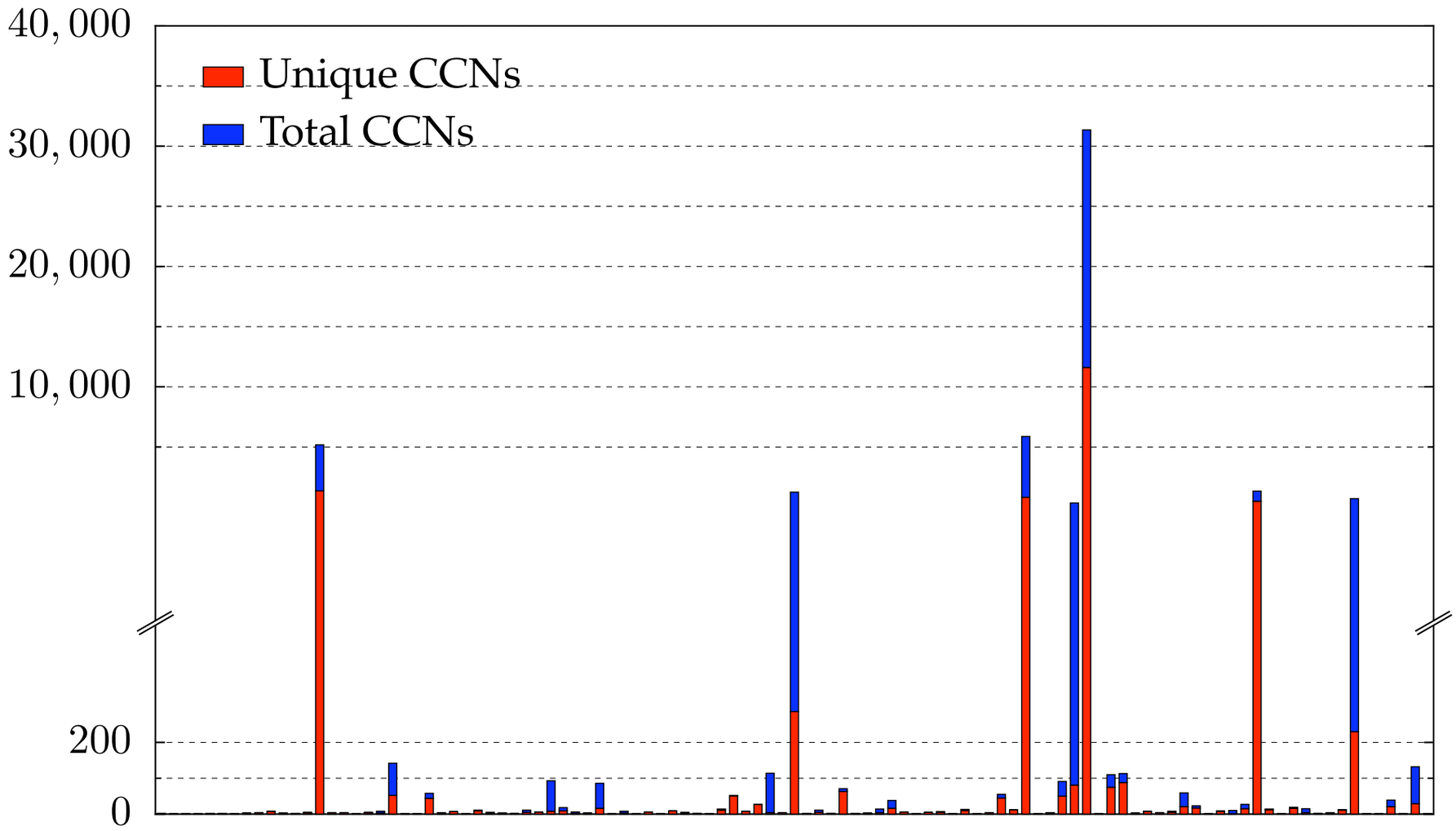
Disk #105:

Test	# pass
pattern	3857
Known prefixes	90
CCV1	43
patterns & histogram	38

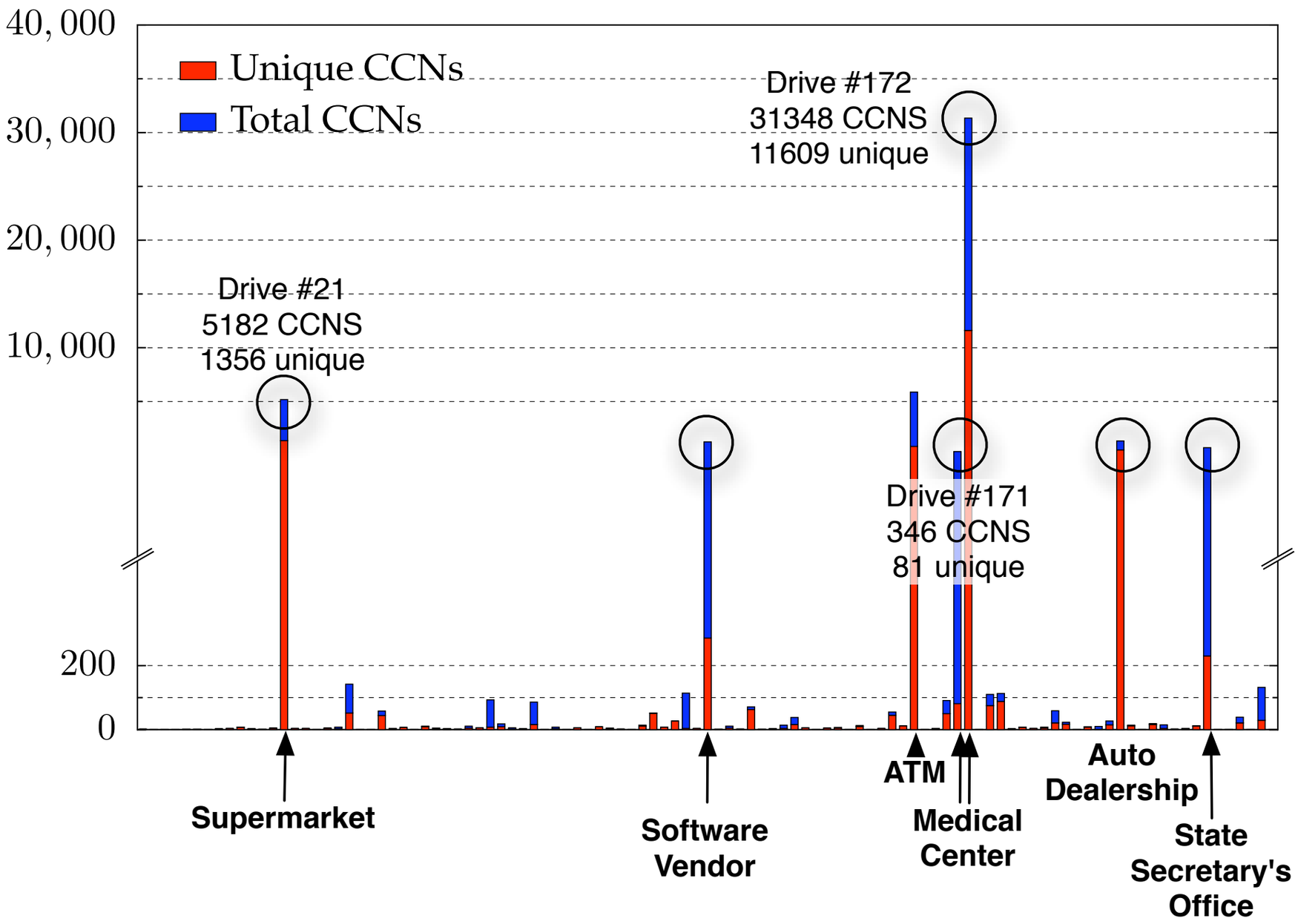
Sample output:

```
13152133     'CHASE NA|5422-4128-3008-3685|
13152440     'DISCOVER|6011-0052-8056-4504|
13152589     .'GE CARD|4055-9000-0378-1959|
13152740     BANK ONE |4332-2213-0038-0832|
13153182     .'NORWEST|4829-0000-4102-9233|
13153332     'SNB CARD|5419-7213-0101-3624|
```

Most drives had just a few, but some had a *lot* of credit card numbers.

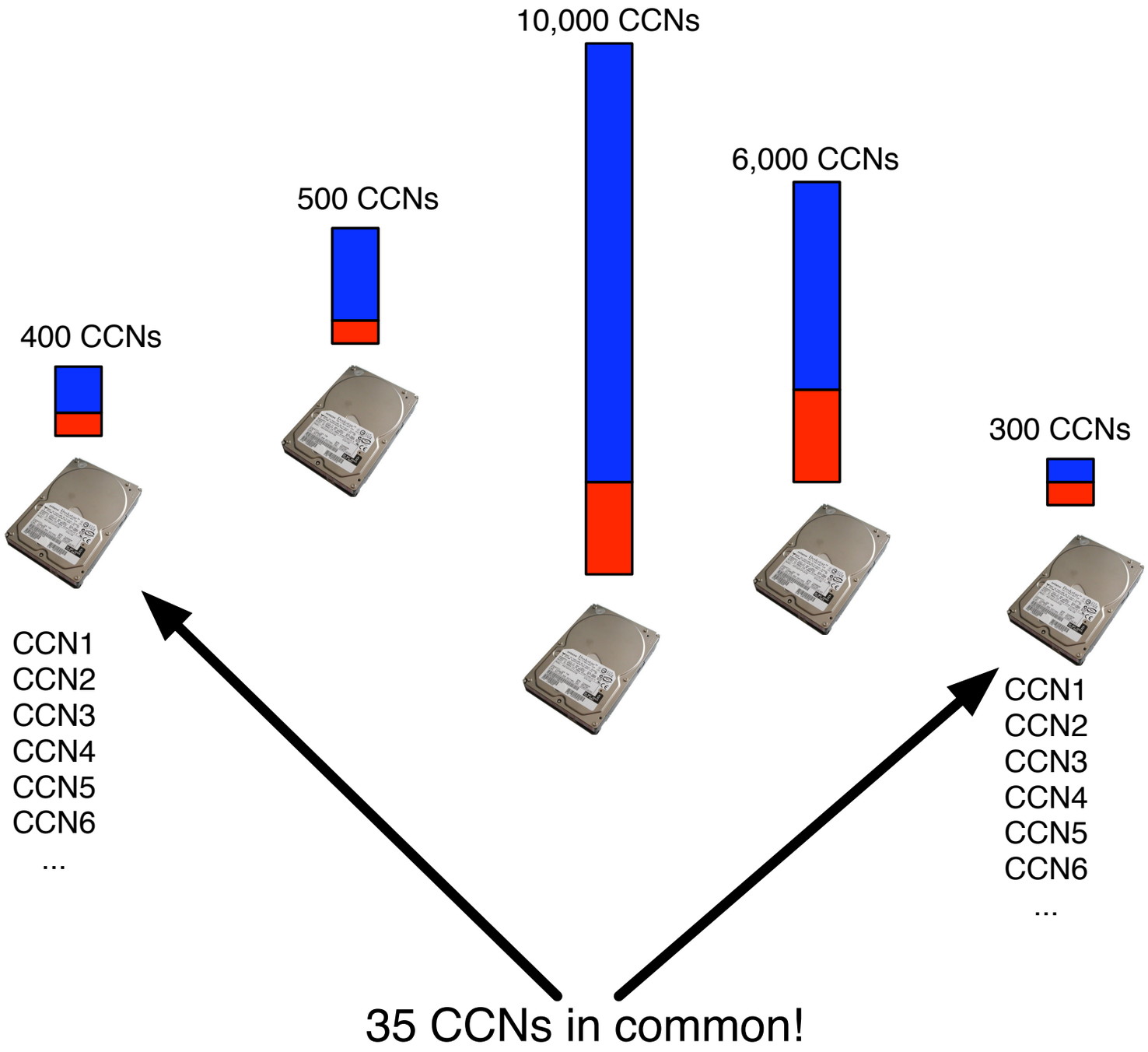


Most drives had just a few, but some had a lot of credit card numbers.



"Design Principles for Software that is Simultaneously Secure and Usable," Garfinkel, MIT PhD Thesis, 2005

What would it mean if two drives had a lot of credit card numbers in common?

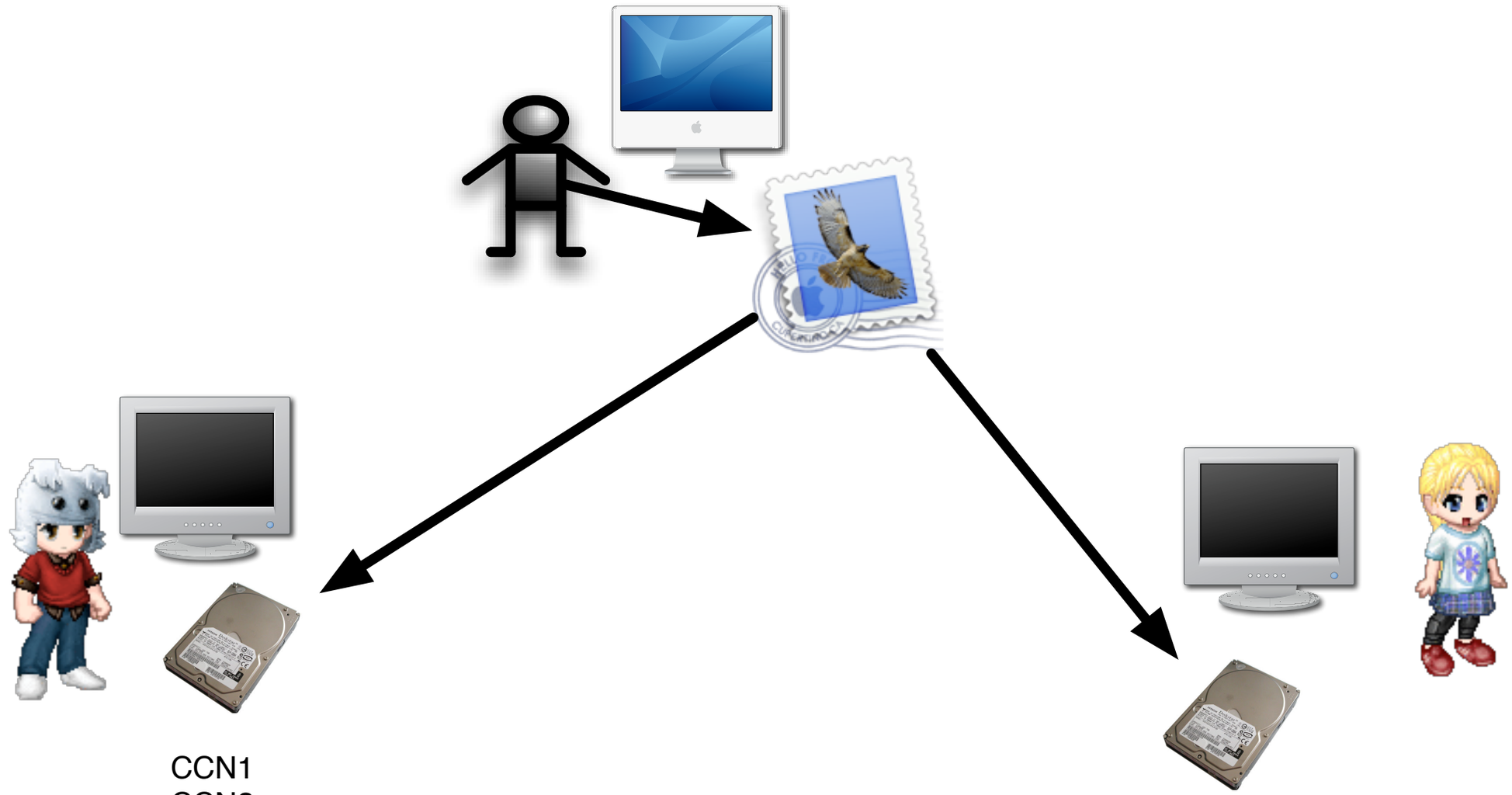


Scenario #1: The owner of one drive sent a message to another drive.

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

Scenario #2:
Both drives received a message from a third party.



CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

Cross Drive Analysis (CDA) computes the correlation matrix of the distinguishing information.

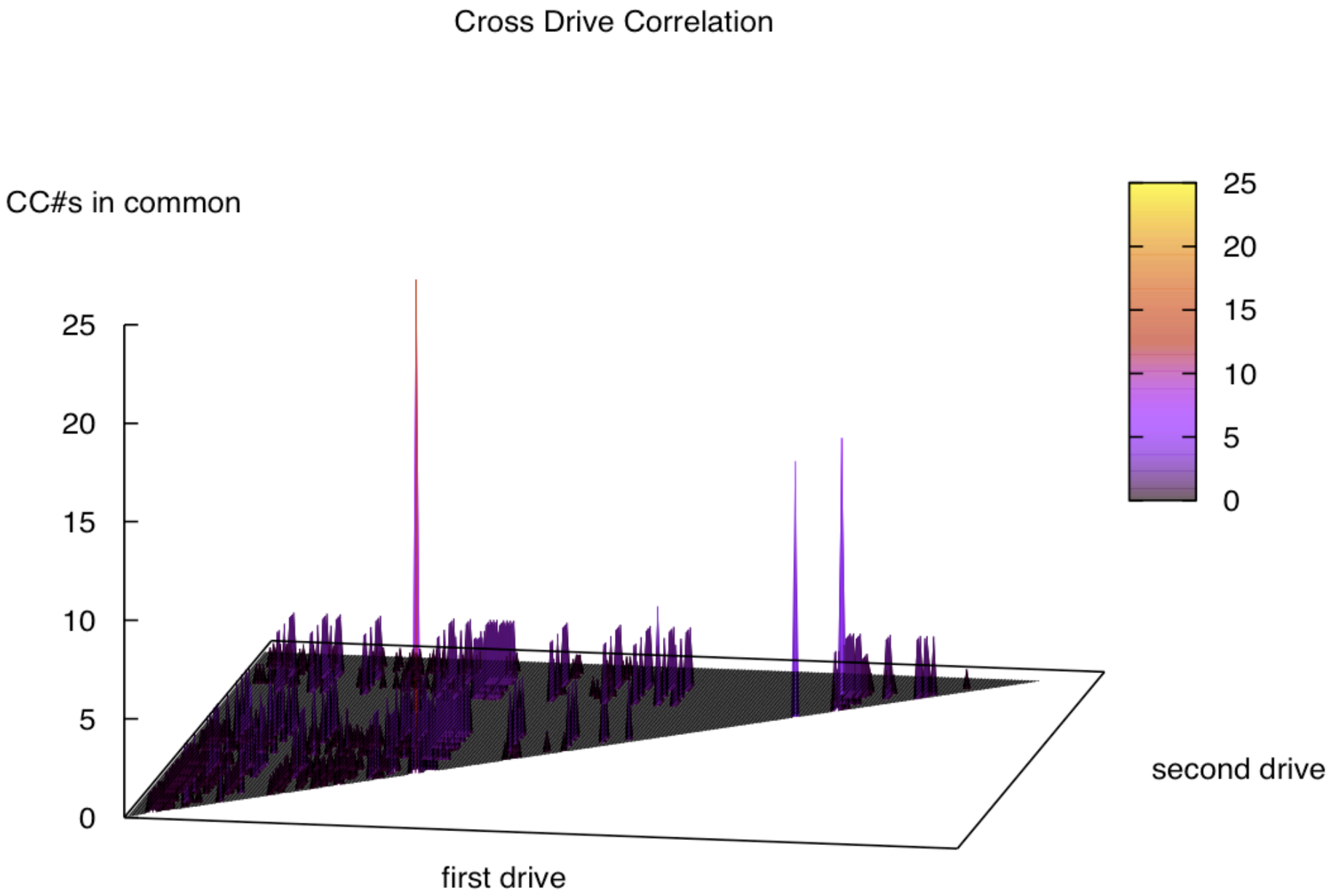
first drive

second drive

2	0	0	4	0
1	1	3	**35**	
0	3	1		
1	0			
4				

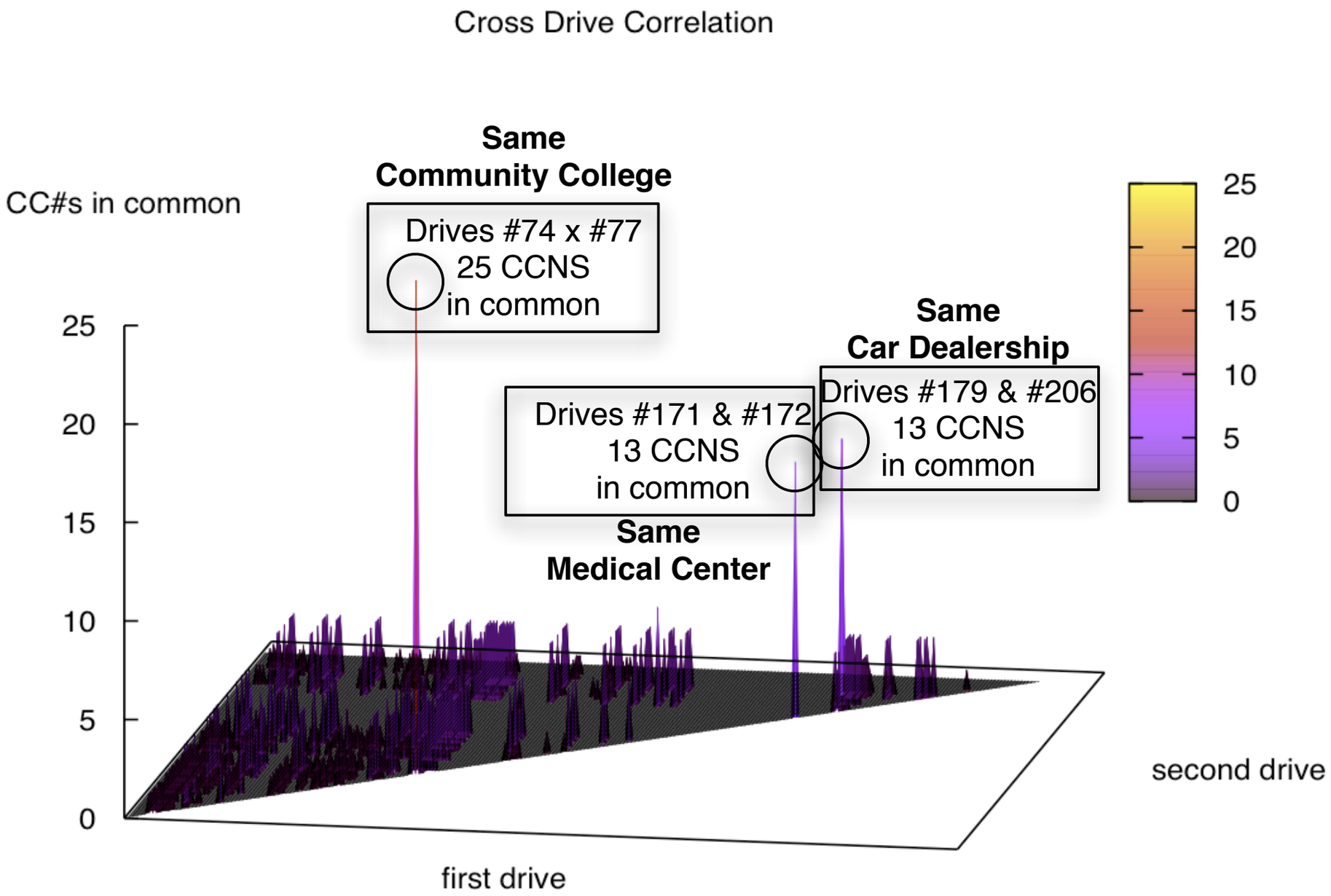
High correlation

This correlation of 250 drives automatically identified media from the same organization.



Garfinkel, S., *Forensic Feature Extraction and Cross-Drive Analysis*,The 6th Annual Digital Forensic Research Workshop Lafayette, Indiana, August 14-16, 2006.

This correlation of 250 drives automatically identified media from the same organization.



Garfinkel, S., *Forensic Feature Extraction and Cross-Drive Analysis*,The 6th Annual Digital Forensic Research Workshop Lafayette, Indiana, August 14-16, 2006.

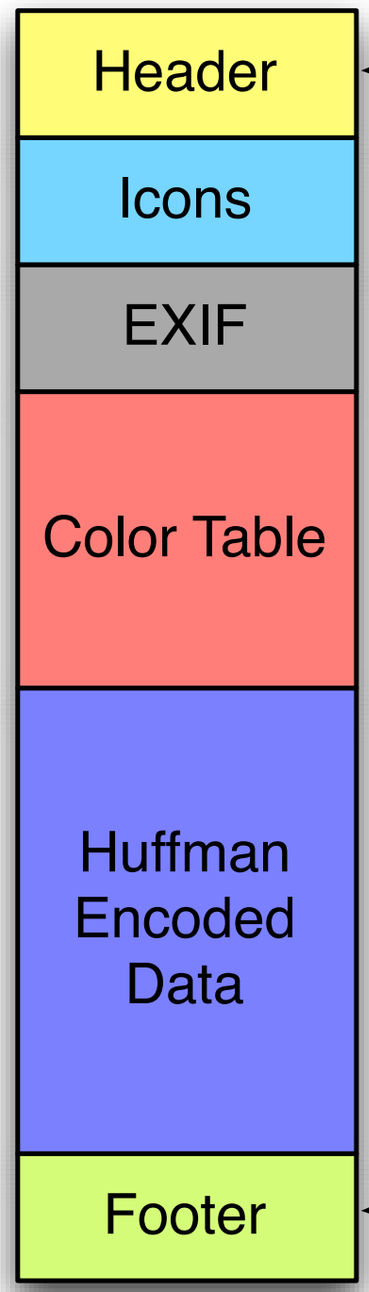
Sector hashing uses fragments of JPEGs as identifiers.

A relatively small JPEG:

Bytes: 31,046

JPEG files have internal structure

Bytes: 31,046



[FF D8 FF E0] or [FF D8 FF E1]

[FF D9]

This JPEG has 61 sectors.

JPEG HEADER @ byte 0

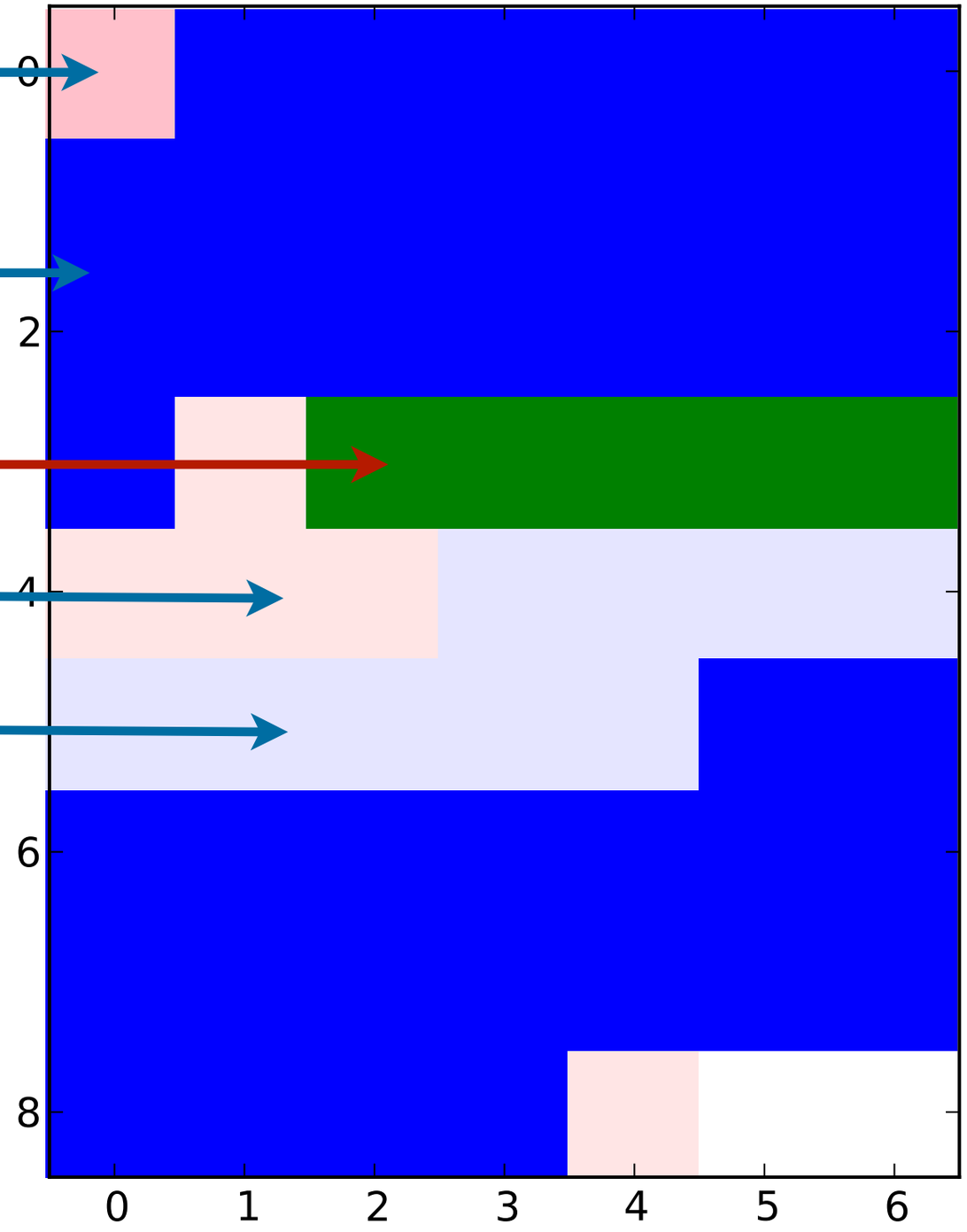
Icons

Exif XML

Exif

Huffman tables

Bytes: 31,046



—Garfinkel, Simson, Vassil Roussev, Alex Nelson and Douglas White, *Using purpose-built functions and block hashes to enable small block and sub-file forensics*, DFRWS 2010, Portland, OR

This JPEG has 61 sectors.
41 of the sectors are “distinct” — not repeated elsewhere.

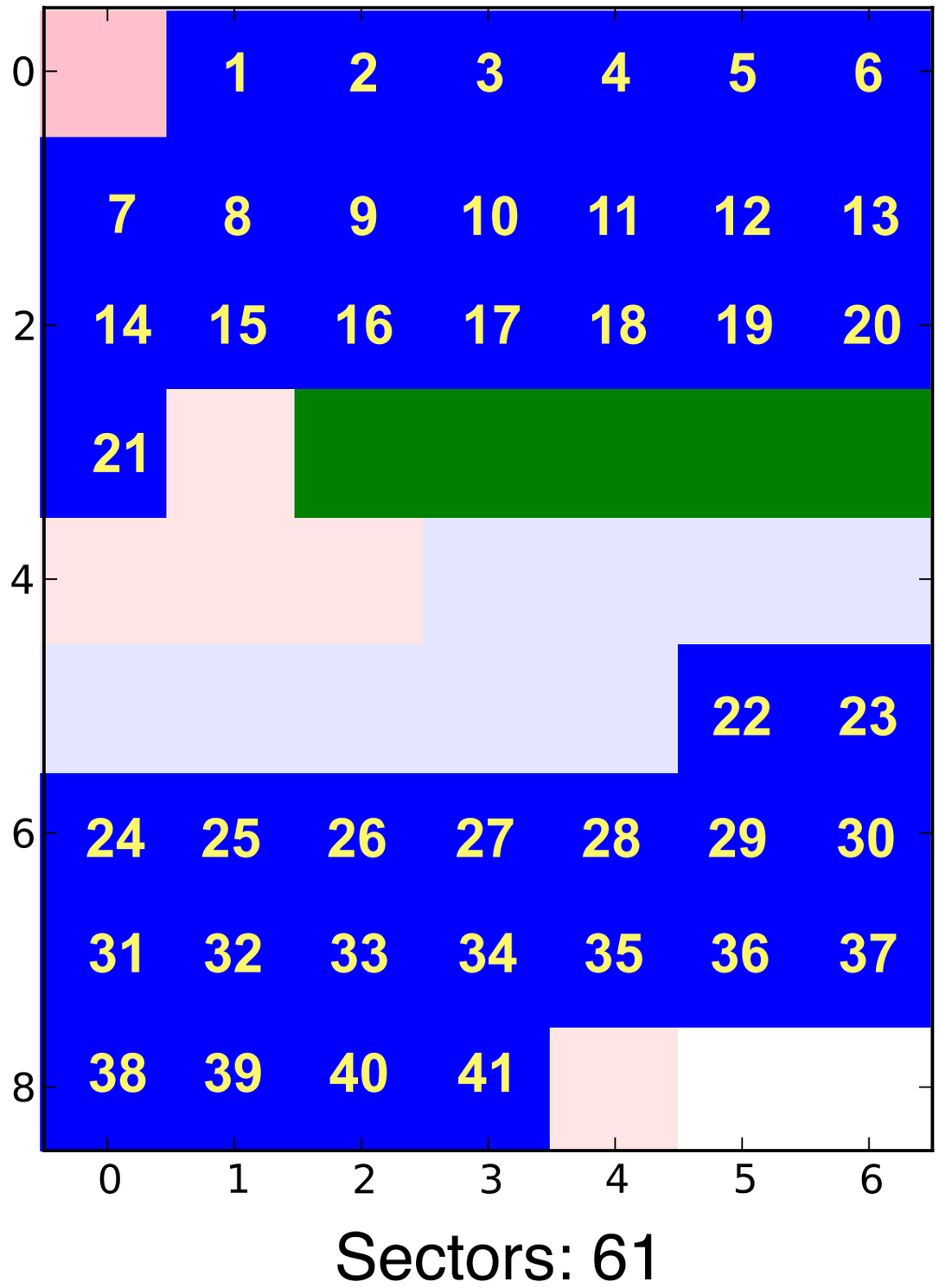
Each BLUE sector is distinct.

- Each has a distinct sequence of bytes

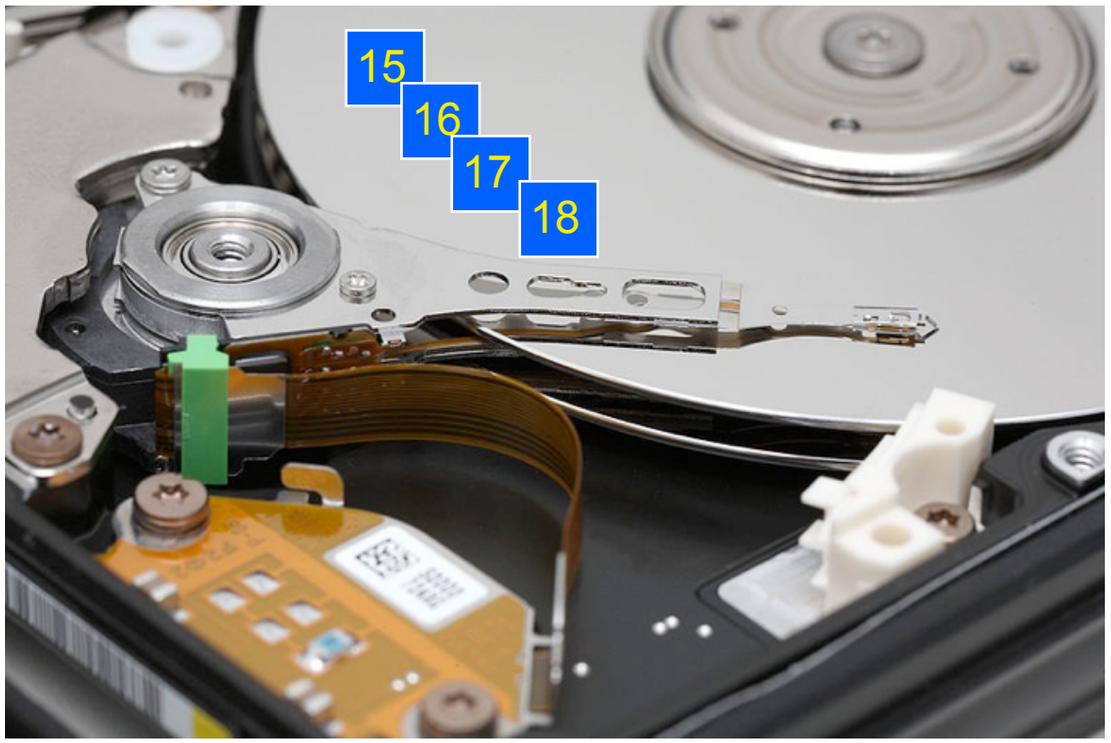
```
dd6a d66f baac 8660 829b b59c 329c 006c
4dbc 5884 5104 3d9b fcae d375 1fab 3ff5
766e 81c0 12f6 b1a0 5bec ab9a f425 9432
02ec bace 23d6 eba0 762b 4b9f 53d0 61de
e003 059c f75c dc9c fdd5 63e2 2696 74af
....
```

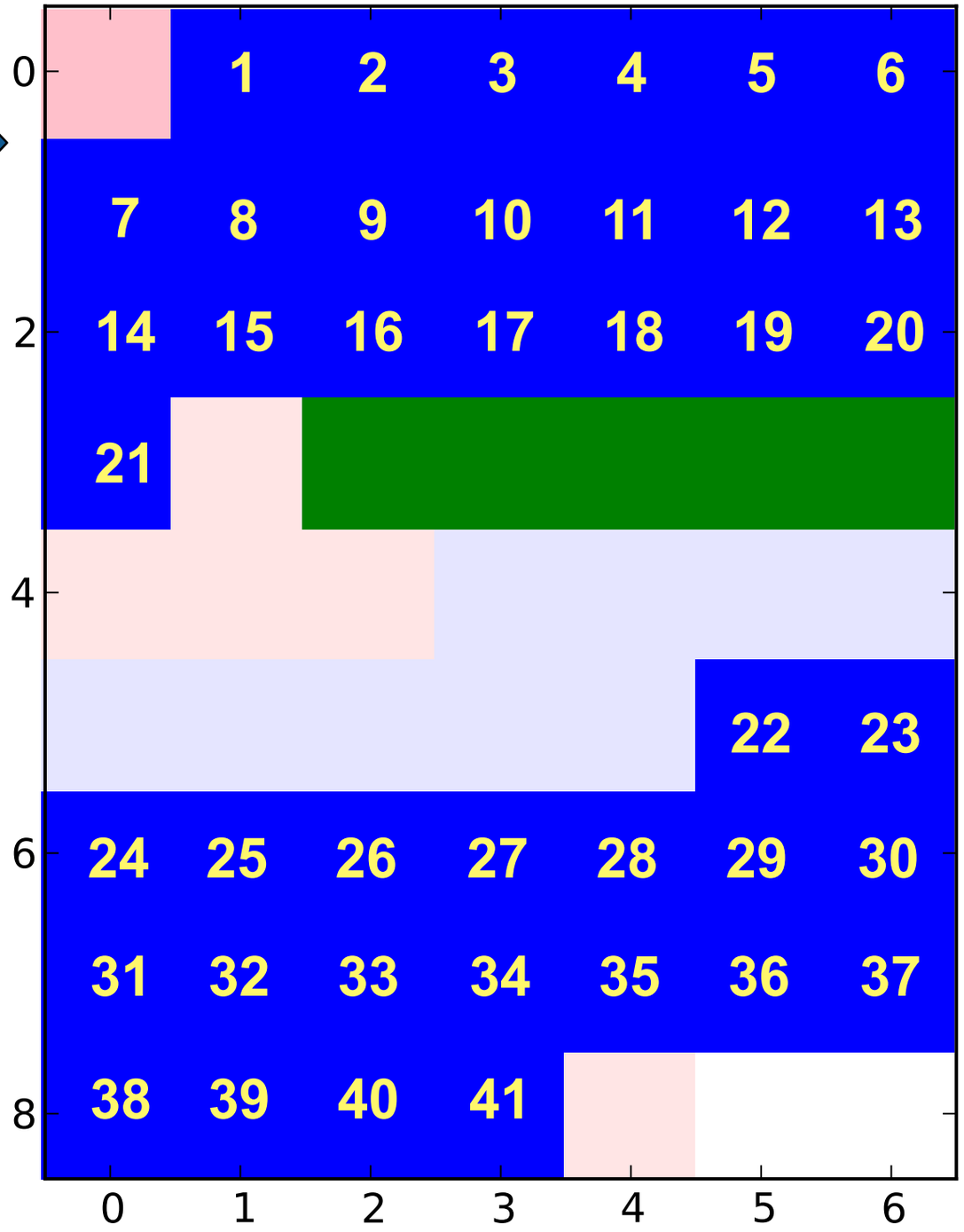
- Each has a distinct MD5 “finger print.”

```
2b00042f7481c7b056c4b410d28f33cf
```

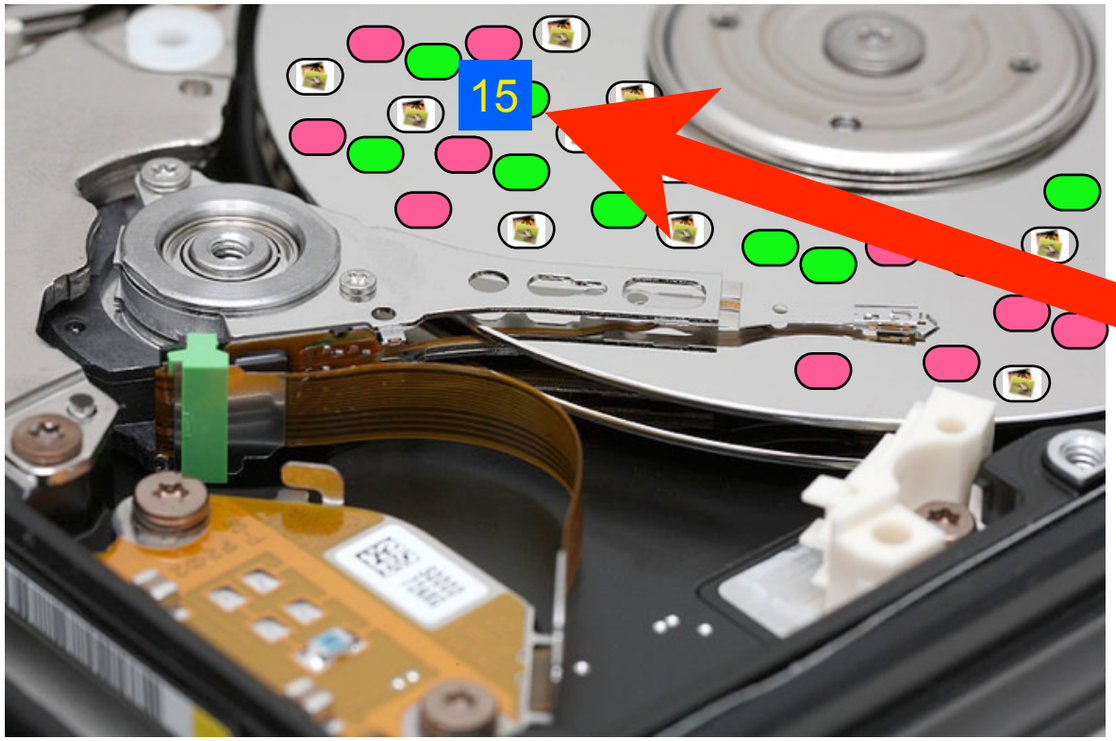


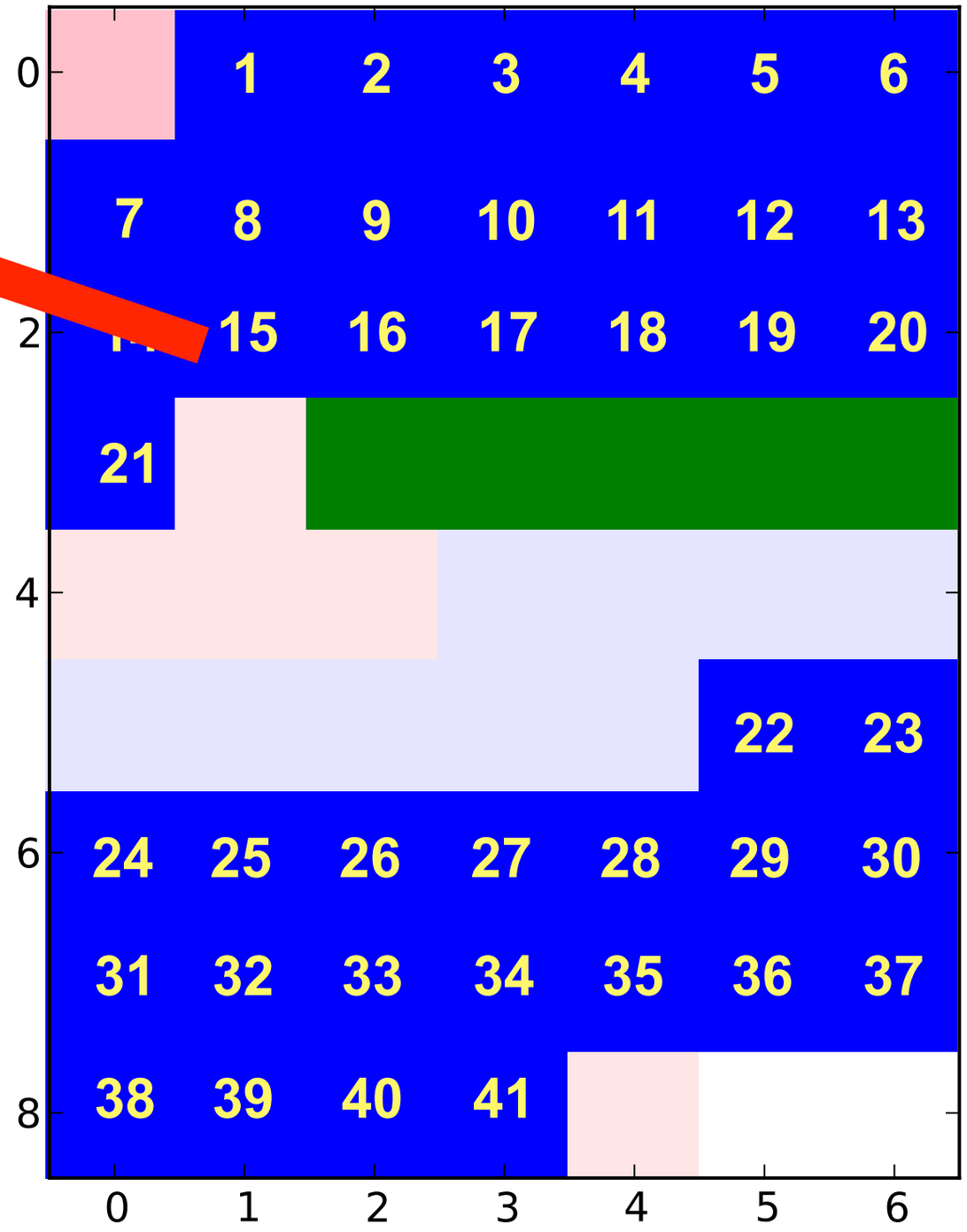
If the image is on a hard drive, each “block” of the image occupies one sector.





The sectors are distinct!
∴ A single sector ➜ the image was probably on the drive.





Two modes of operation:

1. Scan for fragments of deleted files.
2. Use random sampling to scan a 1TB drive in 10 minutes

   p=.99 of finding a sector

Young J., Foster, K., Garfinkel, S., and Fairbanks, K., *Distinct sector hashes for target file detection*, IEEE Computer, December 2012

Law enforcement is a typical application for this technology.

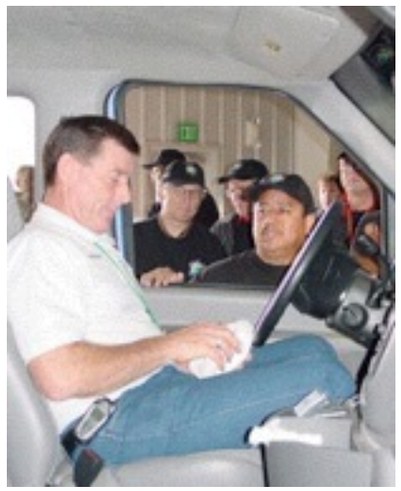
US agents encounter a hard drive at a border crossings...

- Media can be rapidly searched in a way that respects privacy.

  —*5-10 minutes*

- Media can be exhaustively searched.

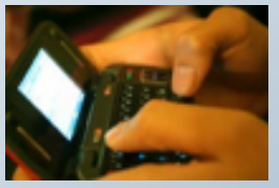
  —*2-3 hours*

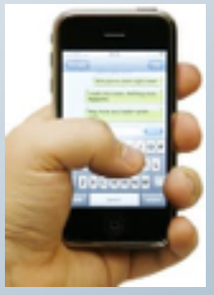


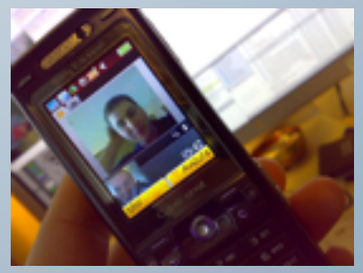
Searches turn up rooms filled with servers….

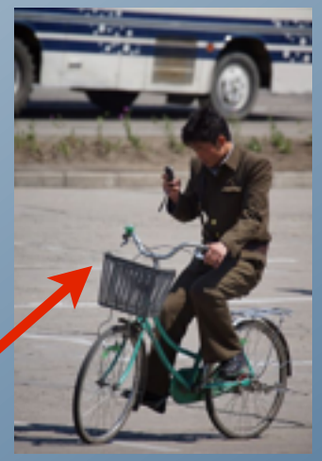
- Systems can be automatically analyzed
- Servers likely to contain evidence can be manually reviewed.
- Connections between servers can be inferred automatically.

Big challenge: tool development and deployment.









Where do we go from here?

For further reading...

***IEEE Computer*, December 2012**

COVER FEATURE

Distinct Sector Hashes for Target File Detection

Joel Young, Kristina Foster, and Simson Garfinkel, *Naval Postgraduate School*
Kevin Fairbanks, *Johns Hopkins University*

Using an alternative approach to traditional file hashing, digital forensic investigators can hash individually sampled subject drives on sector boundaries and then check these hashes against a prebuilt database, making it possible to process raw media without reference to the underlying file system.

Forensic examiners frequently search disk drives, cell phones, and even network flows to determine if specific known content is present. For example, a corporate security officer might examine a suspicious employee's laptop for unauthorized documents; law enforcement officers might search a suspect's home computer for illegal pornography; and network analysts might reconstruct Transmission Control Protocol streams to determine if malware was downloaded. In these and many other cases, examiners typically identify files by computing their cryptographic hash—often with MD5 or SHA1 hash algorithms—and then searching a database for the resulting hash value.

Use of hash values for file identification is pervasive in digital forensics—every popular forensics package has built-in support. One of the most widely used databases is the National Software Reference Library (NSRL) Reference Data Set (RDS). Version 2.36, released in March 2012, contains 25,892,924 distinct file hashes (www.nsrl.nist.gov). Other databases are available to customers of specific companies and to law enforcement organizations.

There are many limitations when using file hashes to identify known content. Because changing just a single bit of a file changes its hash, pornographers, malware authors, and other miscreants can evade detection simply by changing a comma to a period or appending a few random bytes to a file. Likewise, hash-based identification will not work if sections of the file are damaged or otherwise unrecoverable. This is especially a problem when large video files are deleted and the operating system reuses a few sectors for other purposes: most of the video is still present on the drive, but recovered video segments will not appear in a database of file hashes.

SECTOR HASHING

We are developing alternative systems for detecting target files in large disk images using cryptographic hashes on sectors of data rather than entire files. Modern file systems align the start of most files with the beginning of a disk sector. Thus, when a megabyte-sized video is stored on a modern hard drive, the first 4 kibibytes are stored in one disk sector, the second 4 KiBytes are stored in another disk sector, typically the adjacent one, and so on. (In our work, we distinguish between power-of-two-based sizes of digital artifacts, such as kibibytes, and power-of-ten-based sizes, such as kilobytes. See the "Decimal versus Binary Prefixes" sidebar for more details.) Furthermore, by sampling randomly chosen sectors from the drive, it is only necessary to read a tiny fraction of the drive to determine with high probability if a target file is present. This enables rapid triage of drive images.

We compare drive sector hashes to a hash database of fixed-sized file fragments, which we call *blocks*. The terms "sector" and "block" are often used incorrectly as syn-

Published by the IEEE Computer Society

COVER FEATURE

Smartphone Security Challenges

Yong Wang, Kevin Streff, and Sonell Raman, *Dakota State University*

Because of their unique characteristics, smartphones present challenges requiring new business models that offer countermeasures to help ensure their security.

Smartphones are quickly becoming the dominant device for accessing Internet resources. Sales of smartphones overtook PC sales in the global market in Q4 2010.[1] Shipments of smartphones surpassed those of feature phones in Western Europe in Q2 2011.[2] According to a May 2011 Nielsen survey, smartphones outsold feature phones in the US in this same period.[3] Compared to 5.9 billion worldwide mobile phone subscribers, smartphone usage (835 million) is still steadily increasing.[4] IDC predicts smartphone shipments will approach one billion in 2015.[5]

Smartphones offer many more functions than traditional mobile phones. In addition to a preinstalled mobile operating system, such as iOS, Android, or Windows Mobile, most smartphones also typically support carrier networks, Wi-Fi connectivity, and Bluetooth so that users can access the Internet to download and run various third-party applications. Most smartphones support Multimedia Message Service (MMS) and include embedded sensors such as GPS, gyroscopes, and accelerometers, as well as a high-resolution camera, a microphone, and a speaker.

Smartphones' increasing popularity raises many security concerns.[6-8] Their central data management makes them easy targets for hackers. Since the first mobile phone 2011, malware attacks on the Android platform increased 3,325 percent.[9] As the use of smartphones continues its rapid growth, subscribers must be assured that the services they offer are reliable, secure, and trustworthy.

SMARTPHONE THREATS AND ATTACKS

In a smartphone threat model, a malicious user publishes malware disguised as a normal application through an app store or website. Users will unintentionally download the malware to a smartphone, which carries a large amount of sensitive data. After infiltrating a smartphone, the malware attempts to control its resources, collect data, or redirect the smartphone to a premium account or malicious website.

This model divides a smartphone into three layers:

- The *application* layer includes all of the smartphone's apps, such as social networking software, email, text messaging, and synchronization software.
- The *communication* layer includes the carrier networks, Wi-Fi connectivity, Bluetooth network, Micro USB ports, and MicroSD slots. Malware can spread through any of these channels.
- The *resource* layer includes the flash memory, camera, microphone, and sensors within a smartphone. Because smartphones contain sensitive data, malware targets their resources to control them and manipulate data from them.

An attack forms a loop starting with the launch of the malware, moving through the smartphone's application, communication, and resource layers, on to premium ac-

COVER FEATURE

SCADA Systems: Challenges for Forensic Investigators

Irfan Ahmed, *University of New Orleans*
Sebastian Obermeier and Martin Naedele, *ABB Corporate Research*
Golden G. Richard III, *University of New Orleans*

When security incidents occur, several challenges exist for conducting an effective forensic investigation of SCADA systems, which run 24/7 to control and monitor industrial and infrastructure processes.

An industrial automation and control system is a set of devices that regulate the behavior of physical processes. For example, a thermostat is a simple control system that senses the temperature and turns a heater on or off to maintain the temperature at a set point. These systems are used to monitor and control industrial and infrastructure processes such as chemical plant and oil refinery operations, electricity generation and distribution, and water management.

A control system that is spread over a wide area and can supervise its individual components is often called a supervisory control and data acquisition (SCADA) system.[1] However, here we use the term SCADA to refer to all kinds of control systems that share a common key characteristic: they are connected to physical processes and thus need to be continuously available and able to respond within a deterministic time bound.

Early SCADA systems were intended to run as isolated networks, not connected to the Internet, and thus did not require any specific cybersecurity mechanisms. These systems consisted of simple I/O devices that transmitted the signals between master and remote terminal units. In recent years, SCADA systems have evolved to communicate over public IP networks.[2] Some are also connected to a corporate intranet or directly to the Internet to seamlessly integrate SCADA data with external information such as corporate email or weather data.

The integration of SCADA systems within a much wider network brings threats that were unimagined at the time these systems were conceived. During the past decade, vendors, asset owners, and regulators recognized this growing concern and began to address it through new laws and various security mechanisms, processes, and standards.[3]

The discoveries in the wild of Stuxnet in June 2010 and Flame in May 2012 were additional eye-openers for SCADA owners and operators. Stuxnet, the first known malware designed to target automation systems, has infected 50,000 to 100,000 computers worldwide,[4] while Flame is a cyberespionage tool an order of magnitude more sophisticated than Stuxnet.[5]

SCADA ARCHITECTURE

As Figure 1 shows, a typical SCADA system for controlling infrastructures for utilities such as power, gas, oil, or water generally consists of a control center and numerous field sites. The sites are distributed over a wide geographical area and are connected to the control center by different communication media such as satellites, wide

Published by the IEEE Computer Society

For further reading...

“Digital Forensics,” *American Scientist,* Sept-Oct 2013

http://simson.net/clips/academic/2013.AmericanScientist.pdf

Digital Forensics

Finding and preserving evidence of crime in electronic memory requires careful methods as well as technical skill

Simson L. Garfinkel

Since the 1980s, computers have had an increasing role in all aspects of human life—including an involvement in criminal acts. This development has led to the rise of *digital forensics*, the uncovering and in-depth examination of evidence located on all things electronic with memory capacity, including computers, cell phones, and networks. Because of both the scale and the diversity of their domain, digital forensics researchers and practitioners stand at the forefront of some of the most challenging problems in computer science today, including “big data” analysis, natural language processing, data visualizations and cybersecurity.

Compared with traditional forensic science, digital forensics poses significant challenges. Information on a computer system can be changed without a trace; the scale of data that must be analyzed is vast; and the variety of data types is enormous. Just as a traditional forensic investigator must be prepared to analyze any kind of smear or fragment, no matter the source, a digital investigator must be able to make sense of any data that might be found on any device anywhere on the planet—a very difficult proposition.

From its inception, digital forensics has served two different purposes, each with its own difficulties. First, in many cases computers contain evidence of a crime that took place in the physical world. The computer was all but incidental—except because investigators had few tools with which to make sense of his data.

Today personal computers are so ubiquitous that the collection and use of digital evidence has become a common part of many criminal and civil investigations. Suspects in murder cases routinely have their laptops and cell phones examined for corroborating evidence. Corporate litigation is also dominated by electronic discovery of incriminating material.

The second class of digital forensics cases are those in which the crime was inherently one involving computer systems, such as computer hacking. In these instances, investigators are often hampered by the technical sophistication of the systems and the massive amount of evidence to analyze.

Digital forensics is powerful because computer systems are windows into the past. Many systems retain vast quantities of information—either intentionally, in the form of log files and archives, or inadvertently, as a result of software that does not cleanly erase memory and files. As a result, investigators can frequently recover old email messages, chat logs, Google search terms, and other kinds of data that were created weeks, months or even years before. Such contemporaneous records can reveal an individual’s state-of-mind or intent at the time the crime was being committed.

Simson L. Garfinkel earned a Ph.D. in computer science from the Massachusetts Institute of Technology in 2005. He is an associate professor at the Naval Postgraduate School. He holds six U.S. patents for computer-related research and has written 14 books, including Database Nation: The Death of Privacy in the 21st Century (O’Reilley, 2000). Internet: http://simson.net

Information on a computer system can be changed without a trace, the scale of data to be analyzed is vast, and the variety of data types is enormous.

But whereas pre-computer evidence, such as handwritten letters and photographs, could be reproduced and given to attorneys, judges, and juries, computerized evidence requires special handling and analysis. Electronic data are easily changed, damaged or erased if handled improperly. Simply turning on a consumer GPS may cause the device to delete critical evidence. Additionally, computers frequently harbor hidden evidence that may only be revealed when specialized tools are used—for example, a digital camera may appear to have 30 photos, but expert examination may reveal that another 300 deleted photos can be recovered. (When a device “erases” a file, it doesn’t clear the memory space, but rather notes that the space is available; the file may not be really deleted until a new one is written over it.)

Because they can look into the past and uncover hidden data, digital forensics tools are increasingly employed beyond the courtroom. Security professionals routinely use such tools to analyze network intrusions—not to convict the attacker, but to understand how the perpetrator gained access and plug the hole. Data recovery firms rely on similar tools to resurrect files from drives that have been inadvertently formatted or damaged. Forensic tools can also detect the unintentional disclosures of personal information. In 2009 the Inspector General of the U.S. Department of Defense issued a report stating that many hard drives were not properly wiped of data before leaving government service.

Craig Cunningham

Figure 1. West Virginia State Police Digital Forensics Unit. http://www.dailymail.com/News/201108080884

Digital evidence can even be examined to show that something did not happen. Here they are less powerful, for the well-known reason that the absence of evidence is not the evidence of absence. In May 2006 a laptop and external hard drive containing sensitive personal information of 26.5 million veterans and military personnel was stolen from an employee at the Department of Veterans Affairs. After the laptop was recovered in June 2006, forensic investigators analyzed the media and determined that the sensitive files probably had not been viewed.

One way to make such a judgment is by examining the access and modification times associated with each file on the computer’s hard drive. But someone taking advantage of the same forensic techniques could have viewed the laptop files without modifying those timestamps, so the investigators really determined only that the files had not been opened by conventional means.

These examples emphasize that the possibilites of digital forensics are bounded not by technology but by what is practical for people doing the job. Convictions are frequently the measure of success. In many cases there is a considerable gap between what is theoretically possible and what is necessary; even though there may be an intellectual desire to analyze every last byte, there is rarely a reason to do so.

Following Procedures

Digital forensics relies on a kit of tools and techniques that can be applied equally to suspects, victims, and bystanders. A cell phone found on a dead body without identification would almost certainly be subjected to analysis, but so would a phone dropped during a house burglary. How the analysis is performed is therefore more a matter of legal issues than technological ones. As the field has grown, practitioners have tried to create a consistent but flexible approach to performing an investigation, despite such policy variations. Several such *digital forensic models* have been proposed, but most have common elements.

Before data can be analyzed, they are collected from the field (the “scene of the crime”), stabilized, and preserved to create a lasting record. Understanding the inner workings of how computers store data is key to accurately reproducing it. Although digital computers are based entirely on computations involving the binary digits 0 and 1, more commonly known as *bits*, modern computers do most of their work on groups of eight bits called *bytes*. A byte can represent the sequences 00000000, 00000001, 00000010, through 11111111, which corresponds to the decimal numbers 0 through 255

There's still a lot of research to do!

Summer 2013: Analysis of XOR obfuscation in the wild

- 4 interns: 1 Poolesville MD High School student (first author) & 3 West Point cadets

Ecological Studies:

—*Better understanding of what happens inside SQLite3 database files.*

—*Improved exploitation of RFC822/2822 "headers" in Email, Web Servers, etc.*

Identity analytics and disambiguation

- Identify shared accounts or when an email address passes between users.
- Identification of paired (work, home) accounts.

"Big data" and data mining

- Cross-country synchronization of multiple 1PB data sets.

  —*Addressing undetectable read/write errors and silent corruption.*
- Identify hostile insiders with outlier analysis

Visualization and Data Fusion:

- Present complex results in simple, straightforward reports.
- Combine stored data, network data, and Internet-based information.

Contact Information:
Simson L. Garfinkel
slgarfin@nps.edu
http://simson.net/
http://digitalcorpora.org/